CohortVA: A Visual Analytic System for Interactive Exploration of Cohorts based on Historical Data

Wei Zhang, Jason K. Wong, Xumeng Wang, Youcheng Gong, Rongchen Zhu, Kai Liu, Zihan Yan, Siwei Tan, Huamin Qu, Siming Chen, and Wei Chen

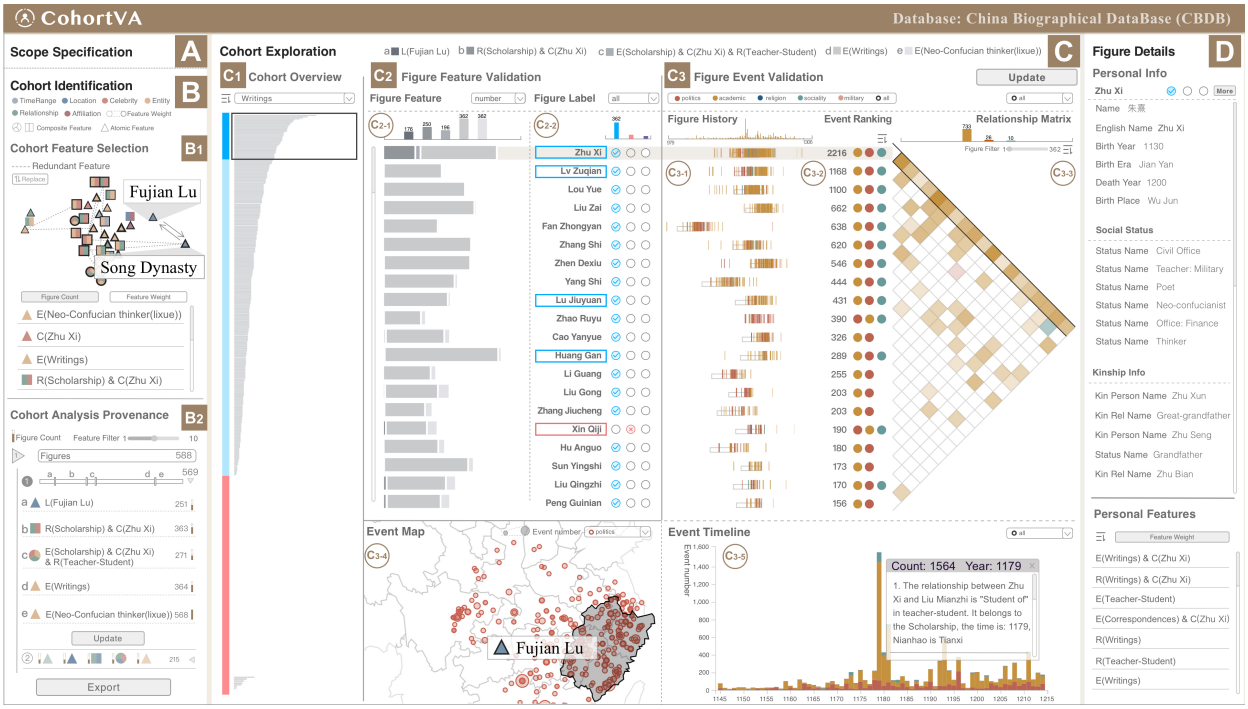

Fig. 1. CohortVA streamlines the iterative cohort analysis workflow and provides visual interpretation for model results. (A) The Scope Specification Component configures the initial research scope. (B) The Cohort Identification Component visually summarizes recommended cohort and feature candidates. (C) The Cohort Exploration Component provides multi-faceted visualizations to validate the selected cohort from multiple perspectives. (D) The Figure Details Component shows the historical figure's profile.

Abstract— In history research, cohort analysis seeks to identify social structures and figure mobilities by studying the group-based behavior of historical figures. Prior works mainly employ automatic data mining approaches, lacking effective visual explanation. In this paper, we present CohortVA, an interactive visual analytic approach that enables historians to incorporate expertise and insight into the iterative exploration process. The kernel of CohortVA is a novel identification model that generates candidate cohorts and constructs cohort features by means of pre-built knowledge graphs constructed from large-scale history databases. We propose a set of coordinated views to illustrate identified cohorts and features coupled with historical events and figure profiles. Two case studies and interviews with historians demonstrate that CohortVA can greatly enhance the capabilities of cohort identifications, figure authentications, and hypothesis generation.

Index Terms—Historical cohort analysis, machine learning, interpretability, visual analytic

- *W. Zhang, Y. Gong, R. Zhu, K. Liu, Z. Yan, S. Tan, and W. Chen are with the State Key Lab of CAD&CG, Zhejiang University. W. Chen is also with the Laboratory of Art and Archaeology Image (Zhejiang University), Ministry of Education, China. E-mail: {zw_yixian, 21951121, zrcccrz, kai.lk, zihanyan, siweitan, chenvis}@zju.edu.cn. Wei Chen is the corresponding author.*
- *J. Wong and H. Qu are with the Hong Kong University of Science and Technology. Email: {kkwongar, huamin}@cse.ust.hk.*
- *X. Wang is with TMCC, CS, Nankai University. E-mail: wangxumeng@nankai.edu.cn.*
- *S. Chen is with Fudan University. E-mail: simingchen@fudan.edu.cn.*

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

1 Introduction

Cohort analysis is a crucial research area in history studies known as *prosopography*, which can inspire the interpretation of the historical process. Here, a *cohort* refers to a group of figures that engage in common activities or have frequent interactions [39]. A prosopographical study typically focuses on one cohort and explores its *concept*, *i.e.*, the set of supplementary features that describes a cohort (*e.g.*, political identities, relationship networks, and social structure) [5, 42]. For instance, based on cohort analysis, Beard [1] revealed that the United States' founding fathers were closely tied with not only leading the American revolution together, but also personal financial interests. This new concept has provoked widespread discussions throughout the U.S. Federal Constitution from a financial perspective. More importantly, it has inspired subsequent works to adopt cohort analysis in suggesting novel concepts and interpretations of well-known cohorts [26, 50].

Recent advancements in digital humanities have greatly relieved historians from manual data collection and labeling. One remarkable progress is that large-scale historical databases, such as China Biographical DataBase [43] and China Government Employee Database-Qing [12], have been carefully built and widely used. Leveraging them in cohort analysis poses new challenges because the data sizes and varieties overwhelm analysts' capabilities [45]. Automatic analysis approaches can alleviate these difficulties yet are far from ideal. First, existing tools focus partially on extracting and analyzing features, but merely support correlation analysis in the context of social structure. For example, Netdraw [7] and Worldmap [23] can only construct basic features from relation networks and geographic locations. Second, most tools are inefficient in integrating human intelligence and supporting iterative exploration. Third, the lack of interpretability prevents historians from effectively verifying obtained results. Historians need additional effort to authenticate results by referring to other documents.

As such, it is highly desirable to integrate domain knowledge and expert hypothesis into the exploration procedure within an intuitive visual interface. For that, we closely worked with historians to observe their behaviors and needs, and identified two challenges: 1) Historians must devote a significant amount of time to synthesizing and cross-checking findings of various figures from a large body of historical literature. It is very time-consuming because figure profiles and features cover various perspectives, such as native place, gender, and changes in official positions. 2) The entire process is repetitive because they may constantly shift the goals. Starting from a familiar targeted group, they might discover interesting patterns and recursively explore and study relevant figures. Considering their exploration process and such obstacles, this is a missed opportunity to provide a visualization-driven solution for efficient analysis and reasoning.

The kernel of our solution, CohortVA, is a cohort identification model that can automatically identify cohorts and their figures. Specifically, given an initial group of interested figures, our model constructs a knowledge graph of related documentary descriptions. It extracts and fuses the common features as the cohort's concept. Based on weakly-supervised learning, our model identifies new cohorts by proposing different cohort concepts as recommendations. The recommended cohorts and their features are visually investigated within the integrated visual interface from multiple perspectives, *e.g.*, location, time, and relationship. In particular, historians can study feature details, reason relations among figures, and progressively refine features and figures of the targeted cohort. Case studies and expert interviews demonstrate that CohortVA frees historians from their heavy workloads and improves the analysis performance. This study makes the following contributions:

- We propose a cohort identification model that utilizes weakly-supervised learning to free historians from manual annotation, querying, and cross-checking.
- We develop a visual analytic approach, enabling historians to study cohorts and concepts interactively.
- We conduct case studies and expert interviews to demonstrate the effectiveness and usefulness of our approach.

2 Related Work

In this section, we review the relevant works in history-oriented visual analysis and visual analytics for cohort studies.

2.1 History-Oriented Visual Analysis

Recently, visual analysis has been widely applied to historical data [8]. History-oriented approaches can be categorized as phenomenon-based and theory-based. A *phenomenon* refers to a recorded historical event, while a *theory* explains why one or more phenomena occurred.

Phenomenon-based research analyzes the phenomena of historical entities such as figures [28, 53], events [14, 24], and cultures [3, 10, 19]. They focus on a few instances in great detail to find hidden patterns and correlations for the phenomenon. Zhang *et al.* [52] contextualized poems from the Chinese Song dynasty with the poets' life stories. The Svoboda Diaries Project [11] uses a person's diaries to recreate the personal experience in Ottoman Iraq. However, they fail to provide a holistic overview for interpreting and generalizing similar phenomena.

Theory-based research deduces coherent explanations from several phenomena. To explore the similarities among social structures, Turchin *et al.* [42] collected and summarized the characteristics (*e.g.*, social scale, economy, and information systems) of 414 societies from 30 regions. Regarding a smaller social unit, GeneaQuilts [4] presents large family trees in an interactive diagonal matrix to study genealogical relationships. Similarly, GenealogyVis [33] explores family structure via correlations between family development and social environment. These works demonstrate the benefits of analyzing social structure from a group perspective. Another popular domain investigates the social mobility aggregated by individual movements. For example, Bol [5] adopted geospatial analysis to study how the Southern Chinese intellectual-social movements spread through the 12^{th} century. Khulusi *et al.* [30] used network analysis and a novel visualization design to define groups interactively for musicians' biography. CareerLens [46] and ACSeeker [47] utilize time-series analysis to explore career trajectories.

In this work, we mainly follow the theory-based approach to assist historians in developing the concepts behind cohorts. We mine features from large-scale multi-dimensional data and fuse them to propose concept candidates. These concepts characterize the cohort and form the basis for understanding the phenomena. We also borrow ideas from phenomenon-based approaches to let users verify the cohorts and contextualize the concepts with detailed historical events.

2.2 Visual Analytics for Cohort Studies

Cohort studies are widely used across various domains, such as medicine [38, 55] and biology [9]. For instance, CoCo [35] integrates statistical and visual analysis for the medical experts to classify and compare cohorts' time series. PhenoStacks [21] simplifies ontological topologies and explores symptom similarities among inter-groups and intra-groups of patients. However, these works focus on analyzing and classifying multiple entities into cohorts with existing and clear definitions. They cannot be directly adopted to historical cohort analysis, which emphasizes the identification of new and vaguely defined cohorts. In addition, they cannot iteratively refine the cohort concepts, which usually only become more evident during the exploratory analysis.

The development of historical databases has contributed significantly to prosopography [6, 27]. Historians widely adopt data visualization tools to explore cohort characteristics from digital records. For example, Gephi [22], Netdraw [7], and Worldmap [23] leverage basic visualization, such as force-directed graphs and choropleth maps, to help historians organize data intuitively. However, these tools fail to integrate multi-dimensional data and features effectively for cohort analysis. Historians need to spend much time cross-validating discoveries about cohorts in the extensive historical literature.

Besides visualization, machine learning techniques have significantly empowered cohort visual analysis in terms of efficiency [16]. For example, Zhao *et al.* [56] used non-binary hierarchical trees and overlapping clustering to shortlist important clusters, reducing the cost of manual selection. However, historians are often confused about the semantic meaning behind the automatic outputs, *i.e.*, the extracted latent features. For instance, Franke *et al.* [20] pointed out that confidence was the first-class attribute of historians for data adoption. They have thus raised awareness of improving interpretability for wider tool adoption.

The closest work to ours is *PK-clustering* [37]. It captures the user's prior knowledge as a set of incomplete clusters, then runs multiple clustering algorithms and visually compares the ensemble results for the user's decisions. It enhances interpretability by keeping human-in-the-loop for each analytical iteration. Although PK-clustering and CohortVA both rely on users' interactions with the interim results, there are still a few key differences compared with our work. First, PK-clustering performs typical clustering tasks on medium-sized datasets with 50-500 entities. In contrast, CohortVA identifies a cluster as one cohort from a large-scale relational database having over 500K historical figures. We proposed a weakly-supervised learning model and interlinked views to address the difference in scale. Second, instead of visualizing the results alone, we provide interpretable features that corroborate the cohort and auxiliary domain information to explain the results. These quickly validate the cohort composition.

3 BACKGROUND

In this section, we introduce the domain background information and outline the requirements obtained through interviewing domain experts. To characterize domain problems and formulate system requirements, we have worked closely with five experienced historians over the past two years. Three of them (H1-H3) are in charge of the China Biographical DataBase (CBDB) [43]. H1 and H2 are committee members, and H3 is a professor leading cohort analysis research projects based on CBDB. The other two (H4 and H5) are Ph.D. students who use CBDB to study the mobility of historical cohorts and the history of the Chinese Song dynasty, respectively. Our collaboration consists of five phases: data acquisition (Sect. 3.1), task analysis (Sect. 3.2), model design (Sect. 4), system design (Sect. 5), and system evaluation (Sect. 6).

3.1 Data Description

This study employs CBDB, a large-scale open-sourced relational database, to study cohorts in Chinese history. CBDB contains enriched biographical records for over 500K historical figures, spanning from the 7^{th} century to the 19^{th} century. The records have been entered and validated by experienced historians. They are multidimensional, covering four main information types:

- *Figures' attributes:* the basic personal information of figures, including birth year, death year, place of birth, place of burial, gender, offices, ethnicity, writing, etc.
- *Figures' relationships:* the social information among figures, including domestic, political, social, and academic relationships, such as colleagues, friendships, kinship, peers, teacher-student, etc.
- *Historical events:* the event description with the time, place, and figure information. As in Fig. 3A2, Zhang Jiuling served as a prefectural aide in 737 for the Jingzhou prefecture, Shannan circuit.
- *Supplementary information:* the detailed description of personal attributes, such as location coordinates, time duration of different dynasties, background knowledge of each dynasty (the governor information and the bureaucratic hierarchy), etc. The amount of information for each figure varies between 5 and 500 records.

3.2 Requirement and Task Analysis

We interviewed our collaborators about their traditional workflow to guide our system design. The traditional historical cohort research follows a mature paradigm [5, 41], including hypothesis formulation, feature summarization, and correlation analysis as shown in Fig. 2A. Our collaborators indicated that the traditional analysis process is energy-exhausting and time-consuming. We summarised the system requirements and design tasks that empower historians with enhanced cohort analysis capabilities, geared toward their challenges.

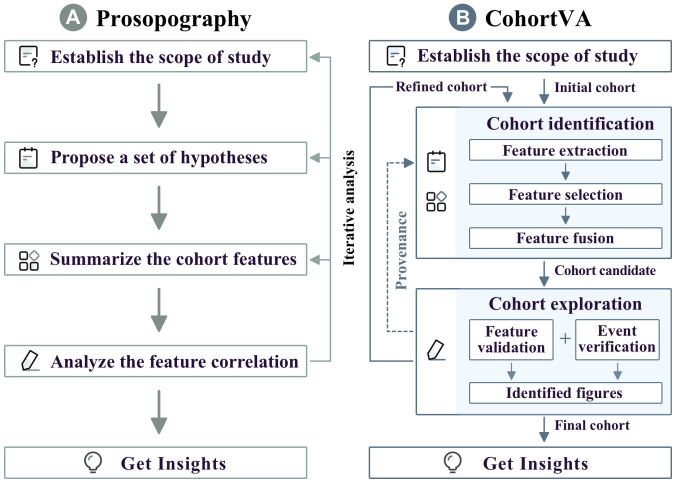

Fig. 2. A comparison between (A) traditional prosopography workflow and (B) proposed workflow applied in CohortVA.

Identifying cohorts from a large-scale historical database. The first step of cohort analysis is to define the scope of the study and formulate corresponding hypotheses. To find the targeted scope, historians need to browse and filter the massive data materials. They then need to perform a heuristic search based on their experience and hypotheses. Verifying a hypothesis sometimes costs researchers a few months to read related materials repeatedly. The system should provide efficient cohort identification based on users' interests.

T1 Specify the initial research scope. Historians' research interests are motivated by different perspectives. For example, supporters of the great man theory would be interested in a few core figures, while those supporting geographical determinism tend to look for location correlations [5]. Moreover, the differences in prior knowledge about cohorts lead to different specifications. We should support different research orientations by specifying the initial groups in multiple ways.

T2 Generate automatic cohort identification results. Traditional prosopography workflow in identifying groups (Fig. 2A) includes browsing massive historical records within the research scope. Automatic cohort identification can alleviate the burden. Since historians might not know everything about a certain dynasty, the automatic method should also account for reasonable ambiguity. Moreover, historians appreciate various cohort candidates complementing cohort analysis from multiple angles.

Providing visual interpretation for cohort identification results. Historians need to cross-check the identified cohorts before using them for further research. The system should provide a set of visual interpretations to assist in result verification.

T3 Validate the concept and features that define a cohort. Historians seek variables of significance to explain a cohort's phenomena. The identified cohort includes the list of figures and the cohort concept. Historians need to validate whether the cohort concept adequately covers the associated figures. Moreover, to find the most suitable concept, historians also want to examine alternative features of the identified cohort.

T4 Verify the cohort from the organized historical event information. H1-H5 emphasize that they always search for additional evidence to verify the results of automated methods. It is necessary to present historians with rich contexts that historians frequently reference, such as geographic locations and social network relationships. They can cross-check the identified cohorts from the detailed historical event information.

T5 Inspect individual figures. Analyzing individual profiles help historians interpret the identified cohort at the most detailed level. All mined features and event evidence can be provided for inspiration. The system should display figure profiles from CBDB with spatial-temporal information and descriptions of social relationships. Directing users to the original sources outside the system should also enhance their trust level in the result.

Supporting iterative cohort analysis. Cohort analysis processes are naturally iterative, and the system should be able to support them.

T6 Adapt to the revising research interests. Historians refine their research scopes iteratively during the analytical process. Due to different research interests, they might disagree on how features are fused and cohorts are composed. For example, Sima Guang and Su Shi are famous writers and political rivals in the Song dynasty. When historians decide to focus on one particular political party, cohorts containing both figures should be discarded, and corresponding features should be remodeled. Our system should update the cohort composition and concept recommendations accordingly, to reflect the revised research interests.

T7 Track analytic provenance. One of the most mundane tasks in prosopography is the inclusion of newly discovered variables. Upon new discoveries, historians need to trace back to previous steps to test what-if scenarios, accompanied by revisiting the same documents repeatedly. We should support revisiting previous progress to compare different cohort identification results.

Table 1. Explanation and extraction models for the six atomic features. Here, descriptions refer to those about the selected figures.

Atomic feature	Explanation	Extraction approach
TimeRange	The time period in which an event happened	Cluster the years occurred in the descriptions with DBSCAN [29]. Then, select the year ranges with more than 30% occurrences as *TimeRange* features.
Location	The location where an event happened	Select the top-three locations in the descriptions as *Location* features.
Affiliation	The government institution where figures held positions	Select the top-three offices in the descriptions as *Affiliation* features.
Relationship	The relationship (*e.g.*, teacher-student) associating two figures	Construct a strongly connected relationship graph based on the descriptions. Then, select the relationships in communities with over five members as *Relationship* features.
Celebrity	The person mostly connected by others	Link any two figures if they appear in the same description. Then, select the figures linked with over 30% of the selected figures as *Celebrity* features.
Entity	The entity, like occasions and writing	Link a figure with an entity if they appear in at least one description. Then, select the entities linked with over 30% of the selected figures as *Entity* features. Only those entities not covered by other features are selected.

4 Cohort Identification Model

The cohort identification model discovers potential cohorts from CBDB based on specified figures and features. As illustrated in Fig. 3A, the model contains four steps: 1) generate the knowledge graph and descriptions, 2) extract common features from the initial figures, 3) select the features by their significance, and 4) fuse the selected features as the cohort concept and filter figures by this concept.

4.1 Knowledge Graph and Description Generation

We preprocessed the raw historical data in CBDB by converting them into a knowledge graph (Fig. 3A1), where the information is stored in nodes (*i.e.*, figures and entities) and edges (*i.e.*, relationships). In CBDB, figures and entities are structured as rows, and their relationships are as foreign keys. For instance, in the [POSTED_TO_OFFICE_DATA] table, an [Posting] event is connected to [Zhang Jiuling] and an office position entity [Prefectural Aide]. The two triplets, (Zhang Jiuling, do, Posting) and (Posting, officeIs, Prefectural Aide), are inferred and inserted into the knowledge graph. We identified 28 node types and 27 edge types, and built a knowledge graph with around 1M nodes and 5M edges.

We adopted the meta-path2vec [18] algorithm to generate descriptions around the figure entities. Collaborating with historians, we summarized 15 description templates that express interpretable and descriptive information, such as politics, occupation, and social relationships. Meta-path2vec leverages the templates and the sequential structure of nodes connected by edges in the knowledge graph. Compared with the conventional random walk method [31], it uniformly generates descriptions regardless of nodes' degrees, avoiding the probability-imbalance issue on different node types. The generated descriptions represent a figure's semantic contexts (see Fig. 3A2 for an example). Lastly, we obtained about 500K figures and 1M descriptions.

4.2 Feature Extraction

We refer to the figure-related characteristics extracted from the descriptions as *features*. Following the conventional cohort study paradigm [5], we categorized figures' characteristics in CBDB into six types of *atomic features*, as shown in Table 1. To represent more complicated contexts, a *composite feature* is generated from multiple atomic features via the *and* logical combination. For instance, the composite feature $[Location(Jingzhou)\ \&\ TimeRange(737)]$ indicates that the figures visited the Jingzhou prefecture in 737.

At the beginning of cohort explorations, users determine a search scope and specify an initial cohort (Fig. 3B1). These initial figures (Fig. 3A1) are the query bases of cohort identification. We generate atomic and composite features from their descriptions. For instance, given the description "Zhang Jiuling served as the prefectural aide in 737 at the Jingzhou prefecture", we use the *TimeRange* model (see Table 1) to extract the time range feature $TimeRange(737)$ and the *Location* model for location feature $Location(Jingzhou)$. The two atomic features jointly form the composite feature $[TimeRange(737)\ \&\ Location(Jingzhou)]$. In addition, since Zhang Jiuling is a renowned scholar-official (further descriptions in Sect. 6.1.2), he is extracted as the feature $Celebrity(Zhang\ Jiuling)$. The three atomic features then generate another composite feature $[Celebrity(Zhang\ Jiuling)\ \&\ TimeRange(737)\ \&\ Location(Jingzhou)]$.

4.3 Feature Selection

Historical records contain numerous descriptions for figures, resulting in a large number of extracted features. However, many features are redundant and insignificant. We adopted the Minimum Redundancy Maximum Relevance algorithm (mRMR) [36] to select appropriate features, which measures statistical dependencies among features.

Given an initial set of n extracted features $\mathbf{F} = \{f_1, f_2, \ldots, f_n\}$, where f_i refers to the i^{th} feature, mRMR selects k features $\mathbf{F}^* = \{f_1^*, f_2^*, \ldots, f_k^*\}$, which are the least redundant and the most significant. It can be individually formulated as follows:

$$min\ R(\mathbf{F}^*) = \frac{1}{|\mathbf{F}^*|^2} \sum_{\substack{f_i^*, f_j^* \in \mathbf{F}^* \\ f_i^* \neq f_j^*}} PMI(f_i^*; f_j^*) \tag{1}$$

$$max\ D(\mathbf{F}^*, \mathbf{F}) = \frac{1}{|\mathbf{F}^*||\mathbf{F}|} \sum_{\substack{f_i^* \in \mathbf{F}^*, f_j \in \mathbf{F} \\ f_i^* \neq f_j}} PMI(f_i^*; f_j) \tag{2}$$

where Eq. 1 aims to select independent features, and Eq. 2 aims to select significant features. The point-wise mutual information (PMI) [15] is applied to measure the redundancy between two features:

$$PMI(f_i, f_j) = log \frac{p(f_i, f_j)}{p(f_i)p(f_j)} \tag{3}$$

where $p(f_i)$ refers to the probability that a figure has the feature f_i, and $p(f_i, f_j)$ denotes the probability that the figure has both f_i and f_j. Statistically, these probabilities can be estimated by the ratio of the figures containing the features. Thus, a higher PMI indicates that the feature pair are more dependent on each other.

We optimize both equations by a genetic algorithm [49], which would yield multiple sub-optimal solutions. Each solution contains k features, where k is defaulted at 5 to balance the model complexity and representation capacity. We call each solution a *feature group*. For every feature f_i^* in one feature group, the two features with the highest PMIs are provided as the *redundant features*.

4.4 Feature Fusion

The selected k features allow each figure to be represented as a k-dim feature vector $\boldsymbol{v} = [v_1, v_2, \ldots, v_k]$, where v_i is the *frequency* of f_i and $v_i = \frac{N_{f_i}}{N_d}$, where N_{f_i} is the number of the figure's descriptions containing feature f_i, and N_d is the total number of the figure's descriptions.

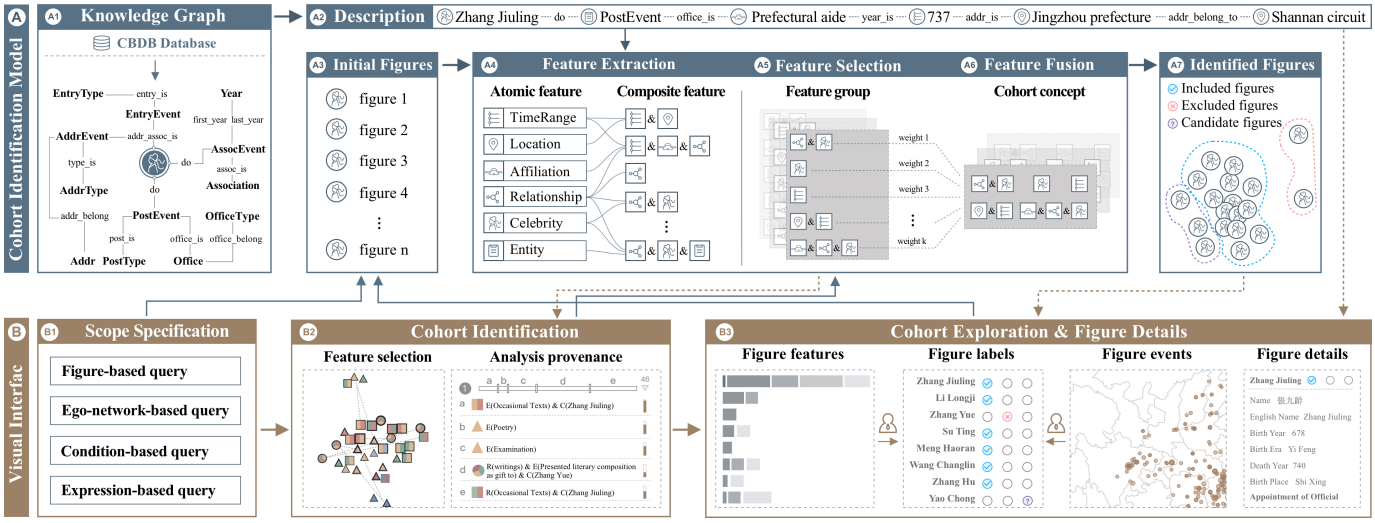

Fig. 3. CohortVA consists of (A) a cohort identification model and (B) a visual interface. We first build (A1) a knowledge graph based on the CBDB and extract (A2) the descriptions. Here, the example from CBDB is translated with [25] to help with the interpretation. Historians specify (A3) the initial figures from (B1) the scope specification component. The figures and extracted descriptions are piped into the model for (A4) feature extraction, (A5) feature selection, and (A6) feature fusion. Then our model automatically identifies (A7) the cohort figures. Generated features and figures are presented in (B2) the cohort identification component, (B3) cohort exploration and figure details component for further interpretation and exploration.

The *concept* of a cohort is defined as the fusion of selected features (Fig. 3A6). It extends the feature group by assigning a fusion weight to each feature. We propose a weakly-supervised classifier to determine whether a figure belongs to the cohort and learn the fusion weights. We calculate the *Cohort Score (CS)* for each figure $\boldsymbol{v}$, having

$$CS(\boldsymbol{v}) = \boldsymbol{w}^T \boldsymbol{v} = \sum_{i=1}^{k} w_i v_i \tag{4}$$

where $\boldsymbol{w}$ is the fusion weights and $\boldsymbol{w} = [w_1, w_2, \ldots, w_k]$. The classifier is implemented as a linear regression model for its high interpretability and the continuous outputs for ranking purposes. The initial figures (Fig. 3A3) are the positive training samples. The stochastic gradient descent (SGD) optimizer is adopted in the learning process.

A figure's membership in a cohort thus depends on its similarity to the cohort's concept. The higher the cohort score, the more likely the figure belongs to the cohort. To account for a possible mismatch from the extracted features and specified figures, every figure will be reassigned a new label regardless of the initial group specification **(T2)**. A figure with a cohort score over 1.0 is included in the cohort. The one whose cohort score is below 1.0 and over 0.5 is viewed as a candidate, and others are excluded. A new cohort is then identified (Fig. 3A7) and presented to historians for visual exploration (Fig. 3B3).

5 Visual Analytic System

We propose CohortVA to present the cohort identification results with explanations and to support iterative exploration.

5.1 Two-stage cohort analysis workflow

CohortVA, a visual analytics system, follows a two-stage cohort analysis workflow as shown in Fig. 2B. To establish the scope of study (**T1**), the *Scope Specification Component* (Fig. 3B1) supports historians in specifying an initial group of figures according to their research interests. The initial group is then piped into the workflow.

Cohort identification stage. Given a defined group, CohortVA produces a series of cohort candidates using the cohort identification model (**T2**). The concepts and features of the identified cohort candidates are validated in the *Cohort Identification Component* (Fig. 3B2) (**T3**). Historians can compare different concepts and adjust the features' weights in the *Cohort Analysis Provenance View*. They can also replace certain features with their redundant features in the *Cohort Feature Selection View*. After validating the concepts, historians select and focus on a cohort candidate for further analysis.

Cohort exploration stage. CohortVA lets historians explore and refine the selected cohort from two perspectives: the cohort concept and included figures. The figures are described by the *Cohort Exploration Component* and the *Figure Details Component* (Fig. 3B3). Historians can cross-check the cohort composition according to historical events **(T4)** and detailed figure descriptions (**T5**). After validation, CohortVA supports historians in excluding figures from the cohort and including related ones in the *Figure Label View*.

Adjustments from either perspective will update the cohort interpretation for the other. Thus, historians can iteratively refine a cohort by piping it back to the Cohort identification stage (**T6**). The analysis process is recorded in the *Cohort Analysis Provenance View* for backtracking and testing what-ifs (**T7**). When the analysis result is satisfactory, historians can export the cohort concept and the list of included figures in CSV format.

5.2 Scope Specification Component

The *Scope Specification Component* (Fig. 1A) provides a control panel for data queries to help historians quickly locate a target figure group.

Enabling flexible group queries. Four means of figure queries are supported: 1) figures-based query, searching figures by name; 2) ego-network-based query, starting with a core figure and expanding with related figures; 3) condition-based query, picking out figures with common descriptions such as year, location, and identity; and 4) expression-based query, allowing skilled historians to write expressions in the feature representation format (Sect. 4.2) to search for figures, which is most flexible but challenging.

5.3 Cohort Identification Component

The *Cohort Identification Component* (Fig. 1B) visualizes the features of the recommended cohorts for historians to select an appropriate cohort and conduct the following exploration. The group features of the search scope and the generated cohort identification schemes are presented in the *Cohort Feature Selection view* and the *Cohort Analysis Provenance view*, respectively.

5.3.1 Cohort Feature Selection View

The *Cohort Feature Selection View* (Fig. 1B1) interprets all identified features from the initial group as the cohort concept. To facilitate concept understanding, this view depicts each feature's importance, similarity, and overall distribution in feature categories. We encode the features in different channels to make a distinction.

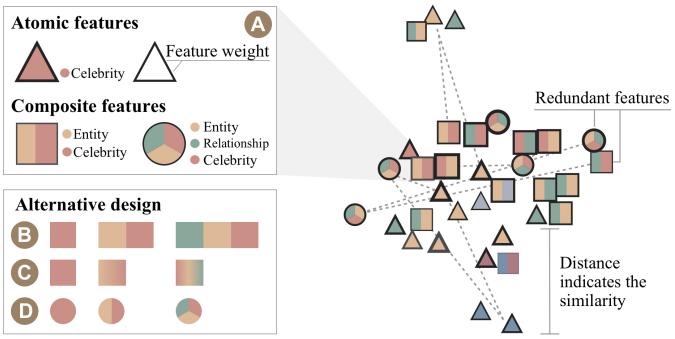

Fig. 4. Cohort Feature Selection View. (A) The encoding scheme for atomic and composite features. (B, C, D) The alternative designs.

Visualizing features in colors, shapes, and thickness. The extracted features could be either atomic or composite (see Sect. 4.2). We use different colors to distinguish the six types of atomic features (*i.e.*, *TimeRange*, *Location*, *Affiliation*, *Relationship*, *Celebrity*, and *Entity*). The composite features combine the colors of their corresponding atomic features. We explicitly encode the number of atomic features by shapes to emphasize the distinction. Specifically, triangles represent atomic features, while squares and circles represent composite features consisting of two and three atomic features, respectively (Fig. 4A). The border's thickness encodes the feature's weight.

Justification. We considered three alternative designs. The first one (Fig. 4B) employs squares to represent atomic features and groups multiple squares to represent composite features. However, composite features with three atomic features occupy too much space and cause visual confusion. We also visualized composite features with gradient colors (Fig. 4C) and color combinations (Fig. 4D) of their atomic features. However, the color differences are too small to be perceived and distinguishable. Therefore, we use two visual channels (*i.e.*, shape and color) to enhance the distinction and perception of features.

Displaying features' similarities. The extracted features are displayed in the force-directed layout. The distance between two features is proportionate to their reciprocal PMI value (see Eq. 3). Therefore, the distance between two features positively correlates with their similarity. If a feature has redundant features, we link them by dashed lines. As well as the graphical representation, a feature list is shown to view features sequentially. In the list, features can be sorted by the count of corresponding figures or the feature's significance (see Eq. 2).

5.3.2 Cohort Analysis Provenance View

The *Cohort Analysis Provenance View* (Fig. 1B2) visualizes an overview of multiple identified cohorts and tracks analysis provenance.

Explaining the cohort concept. The view lists features' descriptions, weights, and the number of related figures). Historians can replace a feature by selecting a redundant feature and clicking the "replace" button in the *Cohort Feature Selection View*. Each feature's fusion weight can be set via the slider on top of the view.

Recording exploring iterations. Iterative cohort analysis processes may yield multiple versions of the cohort. For each version, the view summarizes the number of total figures, changed figures, and cohort features for cohort comparisons across iterations.

5.4 Cohort Exploration Component

To enhance the interpretability of analysis results, the *Cohort Exploration Component* (Fig. 1C) supports historians in validating cohorts from the model and data perspectives.

5.4.1 Cohort Overview

The *Cohort Overview* (Fig. 5A) shows the cohort score (see Eq. 4) and the labeling status of all figures in the selected cohort. Considering data scalability issues, historians need to select a sub-group of figures of interest and check their descriptions in the *Figure Feature Validation View* and the *Figure Event Validation View*.

Summarizing the feature distributions. The view shows the feature distribution of each figure in the selected cohort. The cohort features confirmed by historians in the *Cohort Analysis Provenance-View* are distinguished in this view by different grayscale values. For a figure, the length of each feature represents the frequency multiplied by the corresponding fusion weight (Sect. 4.4). To focus on a certain feature, historians can sort figures according to the feature's value.

Summarizing figure labels. CohortVA employs a multicolored bar to group figures included in the cohort (colored in *blue*), candidate figures (colored in *purple*), and excluded figures (colored in *pink*) on the left of the view. It shows the status of the labeling progress and gives an overview of the labeled figures.

Justification. We tried to encode the figure distribution with colors used in the *Cohort Feature Selection* (Fig. 1B1). However, the multiple colors of composite features did not scale well with the large number of figures, similar to the alternative design in Fig. 4B. The numerous colors distracted historians from other views, and the limited screen space also poses challenges in distinguishing features. Thus, we unified the features and colored them with different shades of gray.

5.4.2 Figure Feature Validation View

The *Figure Feature Validation View* (Fig. 1C2) shows model-related information of the sub-group selected in the *Cohort OverviewView* in a dual column structure. The *Figure Feature View* (Fig. 5B) in the left column illustrates a detailed feature distribution. Besides the zoomed-in figure distribution, a histogram is employed to reflect the number of figures with each feature. The *Figure Label View* (Fig. 5C) in the right column lists the labels of each figure. Historians can modify labels and guide the cohort identification model to update the cohort.

5.4.3 Figure Event Validation View

The *Figure Event Validation View* (Fig. 1C3) visualizes the life experience of each figure by demonstrating five categories of events (*i.e.*., politics, academic, religion, sociality, and military) of concern for historians. We use color encoding to distinguish the five categories. Detailed event descriptions from five perspectives are described below:

- *Category:* Each row in the *Figure History View* (Fig. 5D) shows the events in a historian-selected category of a figure. The events are visualized by a thin bar, of which the horizontal position encodes the time of the event. A row with dense bars indicates that the corresponding figure was recorded significantly in the category. The time spans of all rows are aligned to support event comparisons.
- *Frequency:* The *Event Ranking View* (Fig. 5E) shows the total number of events and the top three categories of events ranked by quantity. It provides an overview of the figure's identity characteristics.
- *Relationships:* The *Relationship Matrix View* (Fig. 5F) employs a 45-degree-rotated matrix to show the relationship among figures. The element in the i^{th} row and j^{th} column represents the events involving the i^{th} and j^{th} figure. The shades of color encode the event quantity. Since the ordering of matrix visualization has an extensive influence on local structure discoveries [44], we adopted the Girvan Newman algorithm [34], a betweenness-based community detection method, to sort the grids and highlight figures with closer relationships.
- *Location:* The *Event Map View* (Fig. 1C3-4) shows the geographic distribution of all events by circles. The circle size encodes the number of events that happen at a location. Clustered circles highlight the region's importance to the cohort.
- *Time:* The *Event Timeline View* (Fig. 1C3-5) shows the temporal distribution of events. If historians choose a specific year, the detailed information of events happening this year will be displayed for further exploration and validation.

5.5 Figure Details Component

The *Figure Details Component* (Fig. 1D) provides historians with detailed descriptions (*i.e.*, background information and hyperlinks to original source) of a selected figure. The figure's identified features are listed in the view for reference. The information helps historians develop an in-depth understanding to make a labeling decision.

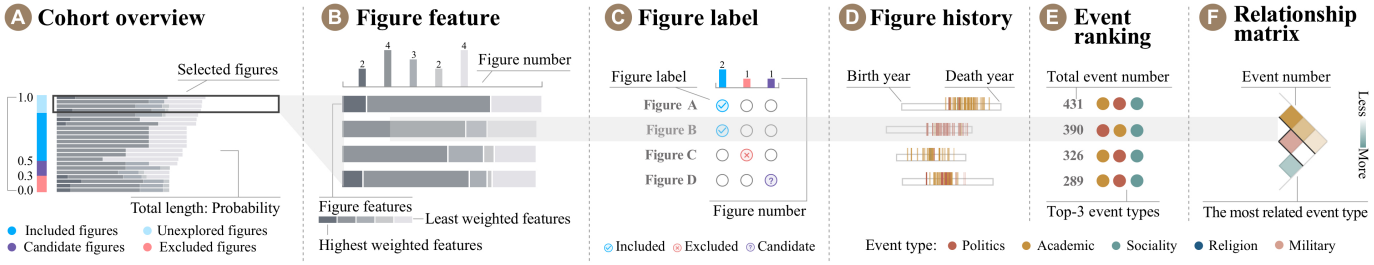

Fig. 5. The Cohort Exploration component. Historians can shortlist figures from (A) the cohort overview. The (B) figure features, (D) figure history, (E) figure events, and (F) figure relationships views provide supporting information for feature- and event-based validation. After cross-checking from both perspectives, historians can label the figures in the (C) figure label view.

6 Evaluation

We conducted two case studies and eight expert interviews to verify the effectiveness and usefulness of our CohortVA.

6.1 Case Studies

We invited the historians mentioned in Sect. 3.2 to explore CohortVA freely according to their research interests and intentions. We encouraged historians to adopt the think-aloud protocol and recorded how they used our system, as described in the following two cases.

6.1.1 Case1: Verify the Neo-Confucianism in Song

The research interest of H4 lies primarily in Neo-Confucianism theory in the Song dynasty. To verify the cohort identified by the traditional prosopography workflow, H4 leveraged CohortVA to explore the Neo-Confucian cohort in the Song dynasty.

Specify the initial cohort of interests. With clear goals in mind, H4 first initialized the figure group through the conditional queries of 'Song' and 'Neo-Confucianists' in the *Scope Specification Component* (**T1**). CohortVA returned 588 figures satisfying the conditions. Then, the cohort identification model (Sect. 4) recommended cohorts in the *Cohort Identification Component* (Fig. 1B) for further analysis (**T2**).

Cohort identification. To select an appropriate cohort candidate, H4 first observed the cohort feature distribution in the *Cohort Feature Selection View* (Fig. 1B1), where several feature clusters appeared. These clusters contained atomic features (*e.g.*, *Writings*, *Neo-Confucian*, *Zhu Xi*) and composite features that are mostly related to *Zhu Xi*, the most famous Neo-Confucianist in the Song dynasty (**T2**). H4 also noticed an outlier feature *Fujian Lu*. H4 indicated that this location feature is significant to the Neo-Confucian cohort because 'Zhu Xi' was born and raised in 'Fujian Lu.' Moreover, after 'Zhu Xi' resigned from the government, he ran colleges to preach Neo-Confucianism in 'Fujian Lu' for forty years. Since 'Zhu Xi' was the core figure, H4 further explored the cohort candidate related to *Zhu Xi* in Fig. 1B2. One of the five identified features of the cohort candidate is *Song Dynasty*, which is too coarse to filter effectively. Thus, H4 replaced it with the redundant feature *Fujian Lu* (**T3**).

Cohort exploration. After determining the cohort features, H4 started to validate the features and refine the figures in the *Cohort Exploration Component* (Fig. 1C). In the *Cohort Overview* (Fig. 1C1), H4 re-sorted the figures according to the feature *Writings*, because authorship was an important characteristic in Neo-Confucianists. Then, H4 screened out the 226 figures with the least cohort scores, and clicked the 'academic' button to observe the writing events that occurred in their lives in the *Figure History View* (Fig. 1C3-1). It turned out that these figures rarely authored writings or participated in academic events, so H4 excluded them from the cohort (**T4**).

As shown in the *Figure Label View* (Fig. 1C2-2), the figures with the highest cohort scores include several representative Neo-Confucianists, such as 'Zhu Xi,' 'Lu Jiuyuan,' 'Lv Zuqian,' and 'Huang Gan.' However, H4 spotted a few included figures that are not Neo-Confucianists when browsing figure descriptions. For instance, the figure 'Xin Qiji' has not contributed to the spread of Neo-Confucianism theory. 'Xin Qiji' was misidentified due to his close political relationship with 'Zhu Xi,' as shown in the *Relationship Matrix View* (Fig. 1C3-3). H4 manually excluded the figure after validating with the events (**T5**). To focus on core Neo-Confucianists and reduce cohort size, H4 also excluded 337 less significant figures with less than one hundred events.

Next, H4 checked the spatio-temporal descriptions of the figures. The *Event Map View* (Fig. 1C3-4) demonstrated that figures in this cohort had recorded events in more than 500 places. Most of the events were clustered in 'Fujian Lu,' which proves the importance of the feature *Fujian Lu*. In the *Event Timeline View* (Fig. 1C3-5), the time feature *1177,1181* attracted H4's attention due to a sudden surge in events. After hovering over the bar corresponding to the time range, H4 found that all these events involve 'Zhu Xi.' For example, 'Zhu Xi' re-built the Bailudong College [48] (one of the four ancient academies in China) in 1179. H4 considered it a milestone event for the spread of Neo-Confucianism.

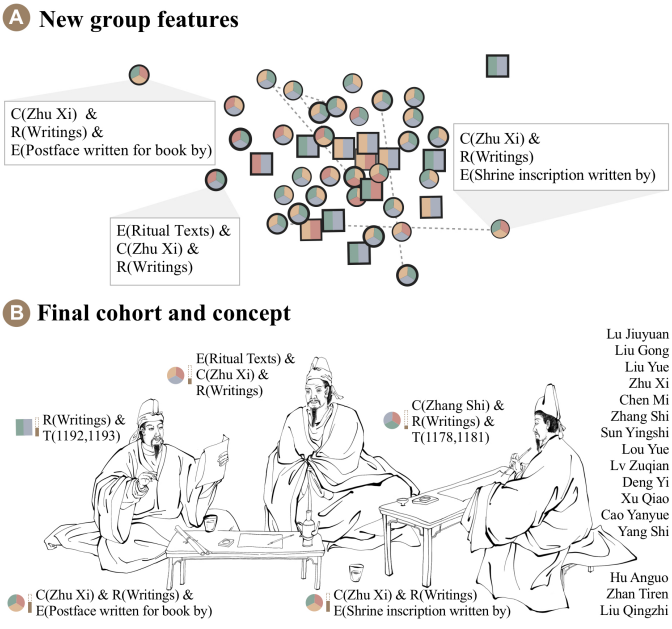

Fig. 6. The identified cohort and concept for Case1. (A) An overview of the new features after the first update. (B) The final cohort and concept.

Iterative exploration. Through the above steps, H4 clicked on the 'update' button in the *Cohort Exploration View* (Fig. 1C), then the system performed the second automatic identification and recommendation in the *Cohort Analysis Provenance View* (Fig. 1B2) (**T6**). New features were selected based on this new cohort. In the *Cohort Feature Selection View* (Fig. 1B1), H4 found more interesting features, such as *Ritual Texts* (the literature embodying Neo-Confucian theories), *Postface written for book by*, and a more precise time feature *1178, 1181* (Fig. 6A). H4 added them into the cohort features (**T6**). H4 checked the recommended figures in the *Cohort Exploration View* and verified that they align with the cohort concept.

Lastly, H4 was satisfied with the cohort having 23 core figures and the more precise concept, as shown in Fig. 6B. Together with some interesting features in previous iterations (**T7**), they were exported in CSV format from the *Cohort Analysis Provenance View* (Fig. 1B2). H4 would further investigate how these figures spread Neo-Confucianism in the Fujian Province. The exported cohort would be cross-checked with other data sources (*e.g.*, local chronicles in *the Intelligent Antiquities Platform* [17]) and verified with other research literature.

This case shows that CohortVA can help historians quickly identify the central figures and important events in the cohort. It provides historians with new perspectives to study the spread of theories.

6.1.2 Case2: Explore the Politics in Poetry

H2 is interested in the history of the Tang Dynasty. In this case, H2 explored the associations around 'Zhang Jiuling,' a famous scholar-official and poet, and the corresponding social influences.

Specify the initial cohort of interests. H2 used the figure-based query to choose figures related to 'Zhang Jiuling' in the *Scope Specification View* (**T1**). 46 figures were selected from 500K historical figures. Then, in the *Cohort Feature Selection View*, CohortVA recommended several cohort candidates whose features were mainly political and literary (**T2**). H2 selected a cohort with the most political and literary features in the *Cohort Analysis Provenance View* for further exploration (**T3**). Besides the feature *Zhang Jiuling*, H2 noticed three identified features of this cohort: 1) *Examination* indicating that the figures had become bureaucrats through the imperial examinations; 2) *Jingzhao* reflecting that this group had frequently been active in 'Jingzhao Fu,' the official designation of the Tang's dynastic capital; 3) *Presented literary composition as a gift to* demonstrating the prevalent literary culture among the cohort (Fig. 7A) (**T3**).

Cross-checking and refining the selected cohort. H2 selected the figures with the highest cohort scores and checked their profiles (**T5**). H2 found that a large proportion of these figures were high-ranking officials. For example, 'Li Longji' and 'Zhang Yue' (Fig. 7B1) were the emperor and the prime minister of the Tang dynasty, respectively. The event distributions in the *Figure History View* showed that the political and academic events were the most frequent event types (**T4**). Combined with the relational information in the matrix (Fig. 7B3), H2 concluded that the tight clique of 'Zhang Jiuling' had been deeply associated with academics and politics, especially the political power center of the Tang Dynasty. H2 removed the figures with less than five academic or political events recorded (Fig. 7B2).

After clicking the 'update' button, the percentage of figures with the feature *Examination* had increased, and a new feature *Poet* had appeared (**T6**). In the *Relationship Matrix View* (Fig. 7B3), most of the interactive events about these figures were related to 'Poetry as a gift,' a more precise description than 'literary composition.' From the analysis result, H2 suspected that poets could bond with officials in the Tang Dynasty by presenting poetry. Favored by the officials, these poets could gain an advantage in the imperial examination for selecting state bureaucrats. For example, 'Zhang Jiuling' ingratiated himself with 'Zhang Yue' and 'Li Longji' by writing poems. When he became the prime minister, other poets started writing poems for him as well. Besides, a new feature *710, 712* appeared after the update. Looking at the *Event Timeline View* (Fig. 7C), H2 found that most figures of this cohort gained government positions during this period, where 96 events were recorded. H2 recalled that the Tang Long coup occurred in 710, after which the regime alternation caused significant bureaucratic alternations. After comparing with other cohorts in previous iterations, H2 decided to include this feature **(T7)**.

Export the exploration results. Finally, H2 exported the exploration results. H2 obtained a refined cohort (consisting of 17 figures) and derived its completed concept as in Fig. 7D. H2 wanted to determine if this cohort was united in their political views. Therefore, he will consult the official documents of the Tang dynasty on *The AiRuSheng Platform* [2] to further explore the influence of this cohort in politics.

This case shows that CohortVA enables historians to obtain a more comprehensive understanding of the cohort and discover hidden connections between features.

6.2 Historian Reviews

To evaluate the effectiveness of CohortVA, we invited eight historians (*i.e.*, three professors (H1-H3) and five PhDs (H4-H8)) to participate in the interview. We collaborated with H1-H5 for two years, as mentioned in Sect. 3. They participated in multiple design iterations of the CohortVA. Three historians (H6-H8) work on historical research on the Tang, Song, and Ming dynasties, respectively, using CohortVA for the first time. H1-H6 are familiar with the CBDB and digital tools (*e.g.*, Gephi and Netdraw) for historical research. H7 and H8 indicated that they mainly employed paper-based historical documents in their daily research. We interviewed three historians (H1-H3) online. Face-to-face interviews were conducted for others (H4-H8). None of the eight participants were co-authors of this manuscript.

Procedure. Initially, each interviewee was asked to fill out a consent form and a demographic questionnaire about their background. Then, we completed the following steps to collect their comments.

- *Training (20 min).* We introduced our motivation, related definitions, the cohort identification model, and visual designs following the visualization introductory description guidelines [51].
- *Freeform Exploration (45min).* We let historians explore our system freely. During the exploration, historians were encouraged to adopt the think-aloud protocol. Their exploration processes were recorded.
- *Interview (20 min).* We asked them to evaluate our system from three aspects, *i.e.*, the effectiveness of the approach, the reliability of the result, and the usability of the system.

Effectiveness of the approach. All historians agreed that their research could benefit from the proposed analysis approach. Five historians (H2, H4-H7) pointed out that CohortVA can significantly improve research efficiency. H5 mentioned that they usually missed significant features and made tremendous efforts to re-identify features after initial explorations. Automatic feature extraction can free them from exhausting concept extractions and allow them to focus on other important tasks, like analyzing the causes and effects. H4 said, "CohortVA acts like a reference book that I will refer to at different research stages." Despite the preference for traditional research methodologies, H8 was interested in using CohortVA to explore unfamiliar areas for efficient explorations and inspiration. H3 affirmed the significance of our work and believed that CohortVA is an appropriate teaching tool for history classes. Besides, H2 suggested that our approach should further support comparing multiple cohorts, which would be one of our future works.

Reliability of the result. Historians must validate the model output before applying the output in historical research. H5 and H7 indicated that the *Cohort Exploration Component* (Fig. 1C) favored them in understanding the extracted cohort features. As mentioned in Sect. 6.1, contextual information helps historians understand why *Fujian Lu* is included in cohort features. H5 and H6 argued that visual designs and interactions facilitate the connection between interpretation and context. H2 and H4 indicated that adding a hyperlink in the *Figure Details Component* (Fig. 1D) was a great function for checking the original material. H3-H8 pointed out the limitation of a single data source and expressed the wish to introduce more data sources, including user-defined data for cross-checking.

Usability of the system. Historians agreed that each view was well-designed. Among all visual forms, historians (H2-H5) considered *Event Map View* (Fig. 1C3-4) to be the easiest to use, believing it has "the most distinct visualization characteristic and the most concise data presentation." H3 suggested providing maps of different historical dynasties to support exploring various research targets. The *Event Timeline View* (Fig. 1C3-5), visualizing events number in each year, was also popular among historians. Although other visualizations, such as the *Cohort Feature Selection View* (Fig. 1B1) and the *Relationship Matrix View* (Fig. 1C3-3), require some learning costs for historians unfamiliar with data visualization and analysis, H2 commented on them as being valuable visualization projects, "digital humanities will inevitably be more complex but powerful to support complicated analytical tasks." H4 particularly preferred swift interaction designs, like the filtering and sorting functions in the *Cohort Overview View* and the tagging function in the *Relationship Matrix View*.

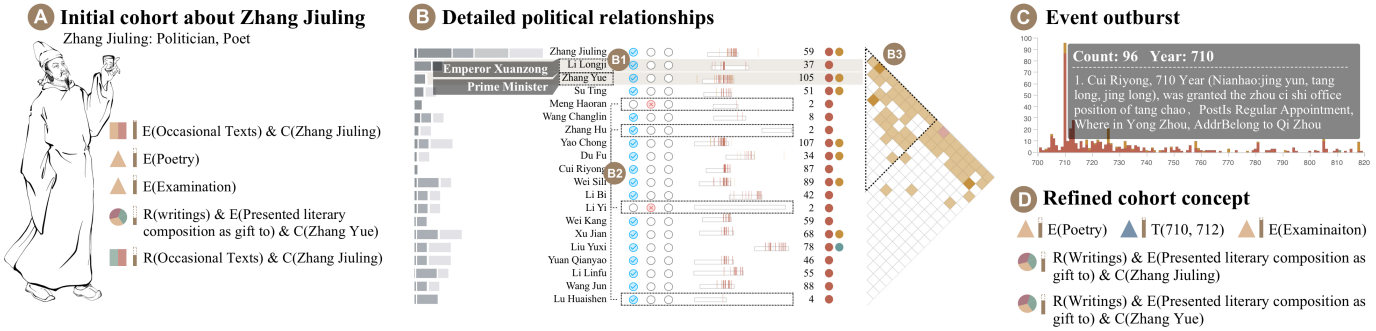

Fig. 7. Exploration process. The initial figures and features of Zhang Jiuling (A) were identified. Then H2 explored the political and academical relationships between these figures (B), and noticed an event outburst (C). After an update, H2 obtained the refined cohort concept (D).

7 Discussion

This section summarizes the lessons learned from working with historians, as well as the limitations and future work of CohortVA. Through collaborations with historians, we have summarized three design implications for digital humanities (**CAD**):

Comprehensibility. Historians prefer easy-to-understand and simple visualizations. Being used to reading a large literature base, they feel more familiar with concrete wordings than abstract visual forms. In our design, we contextualize the generated descriptions with conventional visualizations to simplify the paradigm shifts from reading to perceiving. Moreover, historians are more willing to adopt visualizations that require higher learning costs when presented with clear usage benefits. For example, for the *Relationship Matrix View*, we used a matrix ordering algorithm to cluster similar groups and encoded relationship types with colors. The efficiency in pattern mining has drawn historians' interest in studying. As H3 indicates, "we can quickly discover the types of social relations between figures through the colors on the matrix, which is very novel." However, visual uncertainty should be used with caution because some historians are conservative about probabilistic inference and favor definite evidence.

Authenticity. We have identified four ways to enhance historians' trust and confidence in the method, process, result, and data. 1) Reproducing cohorts that *agree with existing theories* can increase the methods' trustworthiness. It directly verifies functional completeness and correctness. 2) Steering results iteratively give historians more confidence in the process. In CohortVA, historians gradually apply their *prior knowledge and expertise* to annotate and validate the recommended figures. The smaller gaps between iterations help them infer the different results. H3 highly appreciates the provenance tracking function, "which helps me to trace back and adjust my hypotheses for further research." 3) Providing organized information across various perspectives enhances the result's confidence. Historians emphasize *cross-checking* the automatic analysis results to reduce biases and misinformation. 4) Displaying and providing access (*e.g.*, hyperlinks) to the *original data sources* are important features in building trust in the data. Historians value the authenticity of the original textual documents.

Diversity. Since the ancient documents and historical data span thousands of years, special attention should be paid to the diversity in the underlying spatial and temporal contexts. For spatial contexts, map visualizations should be aware of landscape changes. For example, the Yellow River has undergone six major avulsions (*i.e.*, changes in river's course) in history. Ancient times' regional landscapes and geographical characteristics have notable differences from their modern counterparts. Also, the territorial changes in different dynasties would create trouble recognizing and understanding historical events with modern maps. Therefore, we could provide *terrain or historical maps* for reference [40], as suggested by H2.

For temporal contexts, besides the missing data and uncertainty issues [53], we came across the subtle semantic changes in interpreting entities. For instance, the "Prefectural aide" in Fig. 3A2 was represented as "Zhang Shi" in CBDB. From the Qin to Song dynasty, "Zhang Shi" meant the government aide, while it became the administrator from the Yuan dynasty and on [25]. We used composite features to capture these differences implicitly and relied on domain knowledge to spot the difference. H1 suggested that additional knowledge representations could be adopted from *existing theories and structured dictionaries* [25] to provide appropriate contexts and reduce ambiguity.

7.1 Limitations and Future Work

We outline the current limitations to be addressed in future work.

Data source. The current system only employs a single data source. Validation with multiple data sources can further strengthen the interpretability of cohort features. In the future, it should combine with other large-scale databases or self-defined knowledge bases. We could define more description templates for knowledge graph completion, but entity alignment is still a challenging problem.

Cohort comparison. The current system does not provide comprehensive comparisons among cohorts. Analyzing a single cohort limits the research scope and could be biased. Cohort comparison is a complex analysis task we will pursue in our future work.

Quantitative evaluation. Historians agree that our work can improve their trust in machine learning results, but the claim can benefit from a quantitative evaluation. For example, measuring the decision time and compliance with the recommended results provide viable metrics for comparison with other systems.

Generalizability. In this work, we mainly cooperated with historians and developed a bespoke tool for them. Due to different requirements, the interface and specific parametric settings cannot be easily applied to other domains. However, the CohortVA workflow is generalizable to domain-specific tasks targeting closely related groups. By properly defining the features, it can process large text corpora and relational databases for many domains, such as medicine [13], finance [32], and literature analysis [54]. For example, social network relationships and social media posts can replace the history corpus in our system for identifying different social groups among users.

8 Conclusion

In this paper, we present CohortVA, an interactive visual analytic system for historians to identify and explore historical cohorts. Given an initial group of figures, the cohort identification model in CohortVA automatically extracts their common features and identifies potential cohorts to improve the efficiency of historians' research. The visual interface of CohortVA provides various supporting information to help historians cross-check these results, fostering trust in the system and a deeper understanding of cohorts. Two case studies and the historian interviews demonstrate the usefulness and effectiveness of our system. We summarized the lessons learned for developing CohortVA and believe that these design implications will guide system designers in dealing with historical data and working with historians.

Acknowledgments

The authors wish to thank the editors at CBDB for their feedback and contribution to this project. This work was supported by the National Natural Science Foundation of China (No. 62132017, 61972122), Shanghai Municipal Science and Technology General Program (No. 21ZR1403300) and Sailing Program (No. 21YF1402900).

REFERENCES

[1] C. A. Beard. An economic interpretation of the constitution of the united states. *The American Historical Review*, 19(1):162–163, 1913.

[2] Beijing Erudition Digital Research Center. Erudition. http://dh.ersjk.com/jsp/front/prodlist.jsp, 2017. Accessed: 2022-06-15.

[3] A. Benito, R. Therón, A. Losada, E. Wandl-Vogt, and A. Dorn. Exploring lemma interconnections in historical dictionaries. In *Proceedings of Workshop on Visualization for the Digital Humanities*, 2017.

[4] A. Bezerianos, P. Dragicevic, J.-D. Fekete, J. Bae, and B. Watson. Geneaquilts: A system for exploring large genealogies. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1073–1081, 2010.

[5] P. K. Bol. Gis, prosopography and history. *Annals of GIS*, 18(1):3–15, 2012.

[6] P. K. Bol, R. M. Hartwell, M. A. Fuller, et al. China biographical database project (cbdb), 2004.

[7] S. Borgatti. Netdraw software for network visualization, 2002.

[8] A. J. Bradley, M. El-Assady, K. Coles, E. Alexander, M. Chen, C. Collins, S. Jänicke, and D. J. Wrisley. Visualization and the digital humanities:. *IEEE Computer Graphics and Applications*, 38(6):26–38, 2018.

[9] J. Buchmüller, D. Jäckle, E. Cakmak, U. Brandes, and D. A. Keim. Motionrugs: Visualizing collective trends in space and time. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):76–86, 2019.

[10] T. Castermans, H. Hammarström, B. Speckmann, K. Verbeek, and M. Westenberg. Glottovis: Visualizing language endangerment and documentation. In *Proceedings of Workshop on Visualization for the Digital Humanities*, 2017.

[11] A. T. Chen, M. Raza, Y. Gazula, D. Huang, J. Mendoza, and W. G. Andrews. Grounding users in interpretive acts: Lessons learned in the iterative design of a digital collection. In *Proceedings of Workshop on Visualization for the Digital Humanities*, 2019.

[12] 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*, 4(2):431–460, 2020.

[13] F. Cheng, D. Liu, F. Du, Y. Lin, A. Zytek, H. Li, H. Qu, and K. Veeramachaneni. Vbridge: Connecting the dots between features and data to explain healthcare models. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):378–388, 2022. doi: 10.1109/TVCG.2021.3114836

[14] I. Cho, W. Dou, D. X. Wang, E. Sauda, and W. Ribarsky. Vairoma: A visual analytics system for making sense of places, times, and events in roman history. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):210–219, 2016.

[15] K. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[16] J. C. Devi and E. Poovammal. An analysis of overlapping community detection algorithms in social networks. *Procedia Computer Science*, 89:349–358, 2016.

[17] Digital Humanities Research Center, Zhejiang University. Chinese smart ancient book platform. https://csab.zju.edu.cn, 2022. Accessed: 2022-06-15.

[18] Y. Dong, N. V. Chawla, and A. Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 135–144, 2017.

[19] Y. Feng, J. Chen, K. Huang, J. K. Wong, H. Ye, W. Zhang, R. Zhu, X. Luo, and W. Chen. ipoet: interactive painting poetry creation with visual multimodal analysis. *Journal of Visualization*, 25(3):671–685, 2022.

[20] M. Franke, R. Barczok, S. Koch, and D. Weltecke. Confidence as first-class attribute in digital humanities data. In *Proceedings of IEEE Workshop on Visualization for the Digital Humanities*, 2019.

[21] M. Glueck, A. Gvozdik, F. Chevalier, A. Khan, M. Brudno, and D. Wigdor. Phenostacks: Cross-sectional cohort phenotype comparison visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):191–200, 2017.

[22] M. Grandjean. Gephi: Introduction to network analysis and visualisation, 2015.

[23] W. W. Guan, P. K. Bol, B. G. Lewis, M. Bertrand, M. L. Berman, and J. C. Blossom. Worldmap–a geospatial framework for collaborative research. *Annals of GIS*, 18(2):121–134, 2012.

[24] D. Han, G. Parsad, H. Kim, J. Shim, O.-S. Kwon, K. A. Son, J. Lee, I. Cho, and S. Ko. Hisva: a visual analytics system for learning history. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021.

[25] C. O. Hucker. A dictionary of official titles in imperial china. 1985.

[26] E. Iggulden. The "loyalist problem" in the early republic: Naturalization, navigation and the cultural solution. pp. 1783–1850. University of New Hampshire, 2008.

[27] M. Jeffreys, T. Papacostas, J. Bradley, H. Short, and P. Vetch. Prosopography of the byzantine world (pbw). *King's College London*, 2006.

[28] S. Jänicke, J. Focht, and G. Scheuermann. Interactive visual profiling of musicians. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):200–209, 2016.

[29] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady. Dbscan: Past, present and future. In *Proceedings of The Fifth International Conference on the Applications of Digital Information and Web Technologies*, pp. 232–238, 2014.

[30] 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.

[31] G. F. Lawler and V. Limic. *Random walk: a modern introduction*. Cambridge University Press, 2010.

[32] Y. Lin, K. Wong, Y. Wang, R. Zhang, B. Dong, H. Qu, and Q. Zheng. Taxthemis: Interactive mining and exploration of suspicious tax evasion groups. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):849–859, 2021. doi: 10.1109/TVCG.2020.3030370

[33] Y. Liu, S. Dai, C. Wang, Z. Zhou, and H. Qu. Genealogyvis: A system for visual analysis of multidimensional genealogical data. *IEEE Transactions on Human-Machine Systems*, 47(6):873–885, 2017.

[34] M. E. J. N. M. Girvan. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[35] S. Malik, F. Du, M. Monroe, E. Onukwugha, C. Plaisant, and B. Shneiderman. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *Proceedings of the International Conference on Intelligent User Interfaces*, pp. 38–49, 2015.

[36] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[37] A. Pister, P. Buono, J.-D. Fekete, C. Plaisant, and P. Valdivia. Integrating prior knowledge in mixed-initiative social network clustering. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1775–1785, 2020.

[38] R. G. Raidou, O. Casares-Magaz, A. Amirkhanov, V. Moiseenko, L. P. Muren, J. P. Einck, A. Vilanova, and M. E. Gröller. Bladder runner: Visual analytics for the exploration of rt-induced bladder toxicity in a cohort study. In *Computer Graphics Forum*, vol. 37, pp. 205–216. Wiley Online Library, 2018.

[39] N. B. Ryder. *The cohort as a concept in the study of social change*. Springer, 1985.

[40] H. Southall, R. Mostern, and M. L. Berman. On historical gazetteers. *International Journal of Humanities and Arts Computing*, 5(2):127–145, 2011. doi: 10.3366/ijhac.2011.0028

[41] L. Stone. Prosopography. *Daedalus*, pp. 46–79, 1971.

[42] P. Turchin, T. E. Currie, H. Whitehouse, P. François, K. Feeney, D. Mullins, D. Hoyer, C. Collins, S. Grohmann, P. Savage, et al. Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization. *Proceedings of the National Academy of Sciences*, 115(2):E144–E151, 2018.

[43] H. University, A. Sinica, and P. University. China biographical database (cbdb), April 2019.

[44] N. van Beusekom, W. Meulemans, and B. Speckmann. Simultaneous matrix orderings for graph collections. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1–10, 2022. doi: 10.1109/TVCG.2021.3114773

[45] J. Van Zundert. If you build it, will we come? large scale digital infrastructures as a dead end for digital humanities. *Historical Social Research/Historische Sozialforschung*, pp. 165–186, 2012.

[46] Y. Wang, H. Liang, X. Shu, J. Wang, K. Xu, Z. Deng, C. Campbell, B. Chen, Y. Wu, and H. Qu. Interactive visual exploration of longitudinal historical career mobility data. *IEEE Transactions on Visualization and Computer Graphics*, (01):1–1, mar 5555. doi: 10.1109/TVCG.2021.3067200

[47] Y. Wang, T.-Q. Peng, H. Lu, H. Wang, X. Xie, H. Qu, and Y. Wu. Seek for success: A visualization approach for understanding the dynamics of academic careers. *IEEE Transactions on Visualization and Computer*

Graphics, 28(1):475–485, 2022. doi: 10.1109/TVCG.2021.3114790

[48] Z. Weihan. Analyse different characteristics of bailudong college during the north and south song dynasty. *Journal of Suihua University*, 2009.

[49] D. Whitley. A genetic algorithm tutorial. *Statistics and Computing*, 4(2):65–85, 1994.

[50] S. Wilentz. American political histories. *OAH Magazine of History*, 21(2):23–27, 2007.

[51] L. Yang, C. Xiong, J. K. Wong, A. Wu, and H. Qu. Explaining with examples lessons learned from crowdsourced introductory description of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021. doi: 10.1109/TVCG.2021.3128157

[52] W. Zhang, Q. Ma, R. Pan, and W. Chen. Visual storytelling of song ci and the poets in the social-cultural context of song dynasty. *Visual Informatics.*, 5(4):34–40, 2021.

[53] W. Zhang, S. Tan, S. Chen, L. Meng, T. Zhang, R. Zhu, and W. Chen. Visual reasoning for uncertainty in spatio-temporal events of historical figures. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2022.

[54] W. Zhang, S. Tan, K. Liu, L. Shi, S. Chen, and W. Chen. A new perspective on the study of literature (songci): Text correlation and spatio-temporal visual analytics. *Journal of Computer-Aided Design & Computer Graphics*, 31(10):1–10, 2019. doi: 10.3724/SP.J.1089.2019.17970

[55] Z. Zhang, D. Gotz, and A. Perer. Iterative cohort analysis and exploration. *Information Visualization*, 14(4):289–307, 2015.

[56] P. Zhao and C.-Q. Zhang. A new clustering method and its application in social networks. *Pattern Recognition Letters*, 32(15):2109–2118, 2011.