




# State Snapshot Process Discovery on Career Paths of Qing Dynasty Civil Servants

Adam T. Burke (✉)   
Queensland University of  
Technology  
at.burke@qut.edu.au

Sander J.J. Leemans   
RWTH Aachen  
s.leemans@bpm.rwth-aachen.de

Moe T. Wynn   
Queensland University of  
Technology  
m.wynn@qut.edu.au

Cameron D. Campbell  
Hong Kong University of  
Science and Technology  
camcam@ust.hk

**Abstract**—In process mining, computational processing of sequential data allows the discovery and analysis of processes followed by organisations. These can be either explicitly understood processes, captured in documents or rules, or implicit process paths known in more informal or emergent ways. This paper examines a long-lived institution of historical interest, the Qing (1644-1911) Chinese civil service, using data assembled by historians on civil officials during the 19th century. Mapping the promotion process by following paths of officials through civil service postings helps illuminate the everyday operation of the institution and the society around it. Two distinctive features of this data set are that it records states, not events, and careers often include holding multiple concurrent roles. The combination is a poor match for existing process discovery techniques. We describe this structure as a state snapshot log, and present a new discovery technique, the State Snapshot miner, for constructing stochastic Petri net models from such logs. A case study shows its use in analysing promotion paths for elite graduates in the Qing civil service.

## I. INTRODUCTION

To date, process mining [1] has been applied to modern organisations such as businesses and governments, particularly those with large quantities of data available in their IT systems. These techniques allow investigators and managers to understand and improve these organisations, often in an industrial or governmental setting. Process mining can also be used to help understand other organisations and questions beyond management science.

In this paper, we apply process mining to a historic setting. In such a setting, information systems that log detailed millisecond-precision event data are not available. We present a new type of input log that is more likely to be available in historical contexts, a matching stochastic discovery technique, a visualisation of these models, and an application of our approach to historical data of civil servants’ careers in the Qing (1644-1911) dynasty.

Qing dynasty civil service personnel records allow us to understand a historically important institution in Chinese and world history [2], [3]. China has a long tradition of administration by a formal bureaucracy with well-defined procedures for the appointment, review, and promotion, transfer or termination of officials. From as early as the mid-eighteenth century to the end of the Qing in 1911, the government compiled a roster of all regular civil officials, the *jinshenlu* 縉紳錄, every three months. Figure 1 shows a sample page from

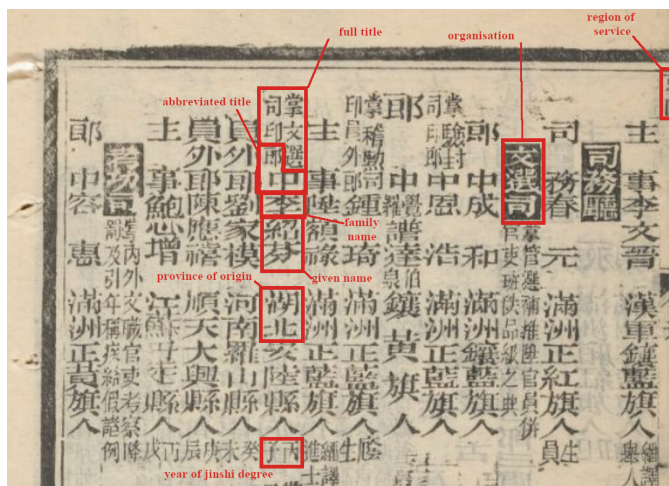


Fig. 1. A civil servant roster *jinshenlu* page from 1897 (volume: Juezhi Quanlan, Guangxu 23, Winter) for a Board Director (郎中), person\_id 188420067500 in the CGED-Q database. Harvard Yenching Library.

this roster for a Board Director (郎中) named Li Shaofen. It includes the official’s name, current roles, qualification, the organisation (e.g. department or bureau) they work in, and other information. An elite official could hold multiple appointments at one time, and these concurrent roles often have mismatched durations. This combination of recording state observations at fixed intervals, and those states holding multiple features from the same feature dimension, is common in other historical sources such as population registers and censuses, due to the expense of recording fine-grained, low latency data, and the usefulness of observation snapshots as a solution to that design problem.

Social science historians have transcribed available editions of the *jinshenlu* from 1760 to 1911 to produce the China Government Employee Database-Qing (CGED-Q) [4], [3]. Their program of historical research using these and related data takes “an exploratory and inductive approach to discover basic facts about the Chinese past that are not readily apparent from traditional approaches in history” [3]. They have used the CGED-Q to study time trends in the composition of civil officials and for case studies on specific groups of officials.

This paper focuses on a process mining research question:

*How may we discover models of processes from records of concurrent states?* This helps advance historical research examining the career paths of Qing dynasty civil servants, as pursued by professional historians. The application of process mining to historical data not produced by a modern information system, and the investigation of the Qing civil service career paths with an computational analysis of aggregate paths, are both novel.

Our solutions are suggested by the structure of this historical data. Firstly, appointments held by a particular official at a particular date are better thought of as a *state* than an *event*. This marks a change from the currently dominant input for process mining: the event log. We instead characterise our input as a *state snapshot log*, formally defined in Section IV-A, where the concepts of cases and traces are retained, but each trace entry is associated with a set of *roles*. Secondly, the output model must represent concurrent appointments. We use a weighted Petri net where places are labelled, rather than transitions. A new discovery algorithm, the *State Snapshot Miner*, is introduced, which constructs these models from state snapshot logs. These techniques are applied to a 74-year window of civil service appointments, particularly the paths of elite graduates. The analysis confirms and complicates the existing understanding of appointments when compared to a normative model based on the official published appointment rules.

In the remainder of this paper, Section II reviews existing work. Section III lays out formal preliminaries. Section IV introduces state snapshot logs and the State Snapshot Miner for process discovery. The case study on Qing civil servant career paths is presented in Section V. Section VI discusses the results, and Section VII concludes.

## II. RELATED WORK

This section surveys existing work in computational history and process mining.

In recent decades, historians and social scientists have begun to use methods that leverage the power of modern computing to describe, model, and analyse change in societies, a trend which can be loosely grouped under the term computational social science [5]. Computational analyses of recurring patterns in parliamentary debate in the French Revolution [6] were one inspiration for this paper. In East Asian history, multi-decade projects to digitise historical sources have made new forms of analysis possible [4], [7]. This research uses the CGED-Q [3], which collects the Qing dynasty civil service records published in the *jinshenlu*. Translations for Qing civil service names have been taken from an established dictionary and related scholarship [8], [9]. Here we focus on a specific subgroup of officials, the holders of the *jinshi* (Metropolitan) examination degree, an elite among officials distinguished by their eligibility for appointment to high office.

Techniques for unsupervised learning of organisational structure from time series data are a form of *organisational mining* [1, p281]. This can include mining organisational models, social network analysis, resource rules or behaviour

profiles, as described by a recent survey [10]. In this sub-field, data attributes associated with operational activities are used to compute models of the organisation at work. In the current study, organisational models are constructed, but the available data is not on operational civil service activities, such as the judgements of local magistrates, or the reports of central government investigators, but on posting and promotion data itself. This allows for the construction of a different type of organisational model, describing career paths as stochastic control-flows.

Life courses, including career paths, are a well established topic in social science research [11]. Sequence analysis methods [12], [11] identify patterns of life stages in a cohort. Sequences are compared using similarity metrics in order to identify clusters or typical paths. In process mining terms, quantitative similarity metrics are calculated by pairwise trace comparison, including using string edit distance, as in the Earth movers' distance measure [13]. Unlike process mining, sequence analysis does not construct a workflow model, or visualise one. Sequence analysis has been used on contemporary and historical data sets. A study of German musicians 1660-1810 [12] identified a number of typical church musician career paths. Like the Qing officials in this paper, musicians could hold simultaneous roles. There, simultaneous roles were excluded; here, our discovery technique accommodates such roles. The sometimes long and complicated career paths of Qing officials have also motivated the development of an interactive browser for CGED-Q data [14], including visualisations of cohort and group mobility; workflow models, as in process mining, would be a complementary addition.

Though the lessons of Roman history for process science have been discussed in scholarly speeches [15], to our knowledge, process mining techniques have not previously been used on datasets from before the advent of modern information systems. Process mining has been applied to modern career paths. An analysis of 35 years of data on employees at a Portugese public organisation [16] found process mining less than insightful, while noting that employees at that organisation rarely changed roles. Direct Follow Graphs (DFGs) were used, and job roles were equated with DFG transitions, on a 585 person data set. University student career paths were used for experiments in declarative process mining [17], using a five-year, 813 student data set from an Italian university. A discovery technique constructed predictive models consisting of both logical rules and associated probabilities, given a conventional event log input.

The State Snapshot Miner that we introduce in this paper creates models in the form of a stochastic Petri net variant. The state-centric reasoning of this algorithm recalls region-based miners [1, p212-222]. There is a growing suite of discovery techniques for stochastic process models, as surveyed in [18]. Recent approaches include labelled stochastic nets with dynamic weights [19] and logical guards [20]. These techniques are designed to work with event logs, whereas the State Snapshot Miner is created to deal with the challenges of state snapshot logs and concurrent roles.

### III. PRELIMINARIES

This section provides formal definitions. Quantifiers and predicates are separated by  $\blacksquare$ , as in  $\exists n \in \mathbb{N} \blacksquare n > 5$ .

The set of sequences over the set  $X$  is denoted  $\text{seq } X$ , and particular sequences denoted by  $\langle a_1, \dots, a_n \rangle$ . The set of values for a sequence is given by  $\text{ran}$ , and concatenation by  $+$ . e.g.  $\text{ran}(\langle 1, 10 \rangle + \langle 7, 1 \rangle) = \{1, 7, 10\}$ .

A multi-set (bag) over a set  $X$  is denoted  $\mathcal{B}(X)$ . The frequency of element  $a$  in bag  $C$  is shown as  $C[a]$ .

A role is an observable element performed in a process. The set of all roles is  $\mathcal{R}$ . In this paper roles are natural language strings describing a civil service appointment, such as ‘‘Senior Examiner’’.

**Definition III.1** (Petri net). A *Petri net* [1, p60], [21] is a directed graph  $(P, T, F)$ , where  $P$  is a set of *places*,  $T$  a set of *transitions*, and  $F \subseteq (P \times T) \cup (T \times P)$  is a flow relation joining particular places and transitions with directed arcs.  $P \cap T = \emptyset$ . A *marking*  $M \in \mathcal{B}(P)$  is a multi-set of places. A transition is *enabled* when every place  $p$  where  $(p, t) \in F$  contains a token.

A transition *fires* by mutating the marking of the net to consume tokens from incoming places and producing tokens for its outgoing places according to  $F$ . An initial marking,  $M_0$ , designates the initial state of the system. For any  $x \in P \cup T$ , predecessor set  $\bullet x = \{y \mid y \mapsto x \in F\}$ , and successor set  $x \bullet = \{y \mid x \mapsto y \in F\}$ .

In order to model both frequencies and concurrent roles, Petri nets are extended with weights and labels in a similar way to stochastic labelled Petri Nets [13], constrained with place capacities [21].

**Definition III.2** (Place-labelled Petri net). A *place-labelled Petri net* is a directed graph  $(P, T, F, W)$ , where  $(P, T, F)$  is a Petri net. Each place has a capacity of one token.  $W : T \rightarrow \mathbb{R}^+$  gives weights for each transition. A *marking*  $M \in \mathbb{P}(P)$  is a set of places with tokens. A transition is *enabled* when every incoming place has a token and, if those incoming tokens are removed, every outgoing place does not have a token. Let  $E$  be this set of enabled transitions for a marking  $M$ .

$$E = \{t \in T \mid \bullet t \subseteq M \wedge (t \bullet \setminus \bullet t) \cap M = \emptyset\}$$

The probability of an enabled transition  $t \in E$  firing is then given by  $\frac{W(t)}{\sum_{t' \in E} W(t')}$ .

The set of all Place-labelled Petri nets is denoted  $\mathcal{N}$ . In this paper, places are roles, i.e.,  $P \subseteq \mathcal{R}$ . An example Place-labelled Petri net is in Figure 2. When a single token is in the Scholar place, the  $t_2$  transition is enabled. When fired, the transition produces a token for Senior Examiner and returns a token for Scholar. The new marking of  $\{\text{Scholar}, \text{Senior Examiner}\}$  does not allow the  $t_2$  transition to be enabled again, as both the Scholar and Senior Examiner output places already hold tokens.

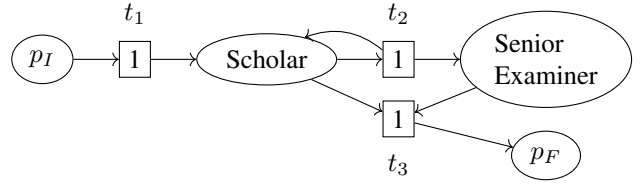


Fig. 2. Example Place-labelled Petri net.

### IV. DISCOVERY WITH THE STATE SNAPSHOT MINER

We introduce a process discovery technique, the State Snapshot Miner, and discuss its implementation, applications and limitations. The miner takes state snapshot logs as an input, which we also introduce formally.

#### A. State Snapshots

A state comprises a case identifier from the set  $\mathcal{U}_{case}$ , a timestamp from the set  $\mathcal{U}_{time}$ , and roles.

**Definition IV.1** (State). A *state*  $s$  is a tuple  $(c, t, z) \in \mathcal{U}_{case} \times \mathcal{U}_{time} \times \mathbb{P}(\mathcal{R})$ .

A state accommodates officials holding one or more bureaucratic positions at any moment in time. In other domains, it could model a network of Internet-Of-Things (IOT) temperature alarms, which can alert both that temperature is too hot and that their battery is running low. Or it could model equities market data, where some exchanges can simultaneously report a price, an auction state, and a halt state for a given stock at a given time.

A sequence of role sets forms a role trace.

**Definition IV.2** (Role trace). Let  $R \subseteq \mathcal{R}$  be a set of roles. A *role trace*  $\sigma \in \text{seq}(\mathbb{P}(R))$  is a sequence of role sets.

The time attribute of the state is used to order the sequence of role sets.

**Definition IV.3** (State snapshot log). A *state snapshot log*  $L$  is a multiset of traces:  $L \in \mathcal{B}(\text{seq}(\mathbb{P}(R)))$ .

The set of all state snapshot logs is  $\mathcal{L}$ . A state snapshot log is constructed from a set of states by creating a trace for each unique case id, ordered by timestamp, and collecting them in a multiset.

As an example, consider the log excerpt in Table I. This holds a sample extract for two officials who were first place, in different years, in the elite palace examination. Note that multiple roles may be held by one official at a particular date.

#### B. Discovery

In general, state snapshot discovery is a function  $\mathcal{L} \rightarrow \mathcal{N}$  transforming a log into a model. We introduce one such algorithm, the State Snapshot Miner. Some straightforward functions for calculating direct follows relations are defined, followed by the algorithm itself.

The function  $dt$  calculates the frequency of a direct follow pair in a sequence.

$$\begin{aligned}
dt &: X \times X \times \text{seq } X \rightarrow \mathbb{N} \\
dt(x, y, \langle \rangle) &= dt(x, y, \langle w \rangle) = 0 \\
dt(x, y, \langle x, y \rangle + t) &= 1 + dt(x, y, \langle y \rangle + t) \\
dt(x, y, \langle w, z \rangle + t) &= dt(x, y, \langle y \rangle + t) \quad \text{given } x \neq w
\end{aligned}$$

The  $d$  function then calculates direct-follow frequencies for a state snapshot log with roles  $R$ , while the  $dh$  function gives frequencies for initial roles and the  $df$  function for final roles.

$$\begin{aligned}
d &: \mathbb{P}(R) \times \mathbb{P}(R) \times \mathcal{L} \rightarrow \mathbb{N} \\
d(x, y, L) &= \sum_{\sigma \in L} dt(x, y, \sigma) L[\sigma] \\
dh, df &: \mathbb{P}(R) \times \mathcal{L} \rightarrow \mathbb{N} \\
dh(x, L) &= \sum_{\sigma_1 \in \{\langle x \rangle + \sigma_2 \in L\}} L[\sigma_1] \\
df(x, L) &= \sum_{\sigma_1 \in \{\sigma_2 + \langle x \rangle \in L\}} L[\sigma_1]
\end{aligned}$$

**Definition IV.4** (State Snapshot Miner). Given a log  $L \in \mathcal{L}$  over roles  $R \subseteq \mathcal{R}$ , the function  $ssm$  outputs a place-labelled Petri net model.

$$\begin{aligned}
ssm(L) &= (P, T, F, W, M_0) \text{ where} \\
p_I &\notin R \wedge M_0 = \{p_I\} \text{ an initial place} \\
p_F &\notin R \wedge p_F \neq p_I \text{ a final place} \\
P &= \{p_I, p_F\} \cup \{a \in A \mid \exists \sigma \in L, s \in \text{ran } \sigma \blacksquare a \in s\} \\
T_I &= \{s \in \mathbb{P}(R) \mid dh(s, L) > 0\} \text{ initial transitions} \\
T_F &= \{s \in \mathbb{P}(R) \mid df(s, L) > 0\} \text{ final transitions} \\
T_R &= \{(s_1, s_2) \in \mathbb{P}(R) \times \mathbb{P}(R) \mid d(s_1, s_2, L) > 0\} \\
T &= T_I \cup T_F \cup T_R \\
F_I &= \{p_I \mapsto t \mid t \in T_I\} \cup \{t \mapsto p \mid t \in T_I \wedge p \in t\} \\
F_F &= \{t \mapsto p_F \mid t \in T_F\} \cup \{p \mapsto t \mid t \in T_F \wedge p \in t\} \\
F_R &= \{p \mapsto (s_1, s_2) \mid (s_1, s_2) \in T_R \wedge p \in s_1\} \\
&\quad \cup \{(s_1, s_2) \mapsto p \mid (s_1, s_2) \in T_R \wedge p \in s_2\} \\
F &= F_I \cup F_F \cup F_R \text{ arcs follow role set pairs} \\
W &= \begin{cases} dh(t) & \text{if } t \in T_I \\ df(t) & \text{if } t \in T_F \\ d(t) & \text{if } t \in T_R \end{cases}
\end{aligned}$$

The miner works as follows. Each role becomes a place. Initial and final places are added to these. Transitions correspond to either sets of roles (for initial and final transitions), or pairs of sets of roles. Flows then join transitions going from one set of roles to another, as indicated by the evidence in the state snapshot log. There is a separate transition for each distinct combination of input and output role sets, and ones to join the initial and final places, for starting and final role sets. The weight of transitions reflects how often this distinct combination occurs. Time intervals where no role is reported

TABLE I  
EXAMPLE STATE LOG EXCERPT,  $L_E$ , FROM CGED-Q DATA.

person id	name	date (y-m)	roles
188330054900	陳冕 Chen Mian	1883-06	{ 修撰 Senior Compiler }
189230033500	劉福姚 Liu Futiao	1892-09	{ 修撰 Senior Compiler }
189230033500	劉福姚 Liu Futiao	1893-03	{ 修撰 Senior Compiler, 正主考 Chief Examiner }
189230033500	劉福姚 Liu Futiao	1894-03	{ 修撰 Senior Compiler }
189230033500	劉福姚 Liu Futiao	1897-03	{ 修撰 Senior Compiler, 副考官 Vice Examiner }

for a given case are treated as continuations of the previous model state (marking), unless they are after the end of the sequence (terminal).

Applying the State Snapshot Miner to  $L_E$  in Table I yields the model in Figure 3.

The time complexity of the State Snapshot Miner is linear in the number of state snapshots and the number of roles. Each state snapshot must be considered in evaluating initial, adjacent, and final role sets. There are  $2^{|R|}$  possible role sets in a log, and so  $(2^{|R|})^2$  possible role set combinations in the worst case. At each iteration through a trace entry, some data structure indexing role set combinations must be updated with frequencies. This data structure (say, a hashtable or tree) can be indexed in log time. The complexity is then

$$\begin{aligned}
O(\|L\| \cdot \log((2^{|R|})^2)) &= O(\|L\| \cdot \log 2^{2|R|}) \\
&= O(\|L\| \cdot 2|R| \log 2) \\
&= O(\|L\| \cdot |R|) \\
&\quad \text{as } 2 \cdot \log 2 \text{ is a constant}
\end{aligned}$$

To allow managing the complexity of output model visualisations by abstracting away from infrequent paths, noise reduction is added to this algorithm by pruning transitions below a weight threshold, and the arcs that connect them.

### C. Implementation and Visualisation

The State Snapshot miner has been implemented in Python<sup>1</sup>. Petri net models are exportable to PNML or as pm4py objects. It includes optional noise reduction by transition weight, taking

<sup>1</sup>Source and sample data are at <https://github.com/adamburkegh/statesnap-miner>.

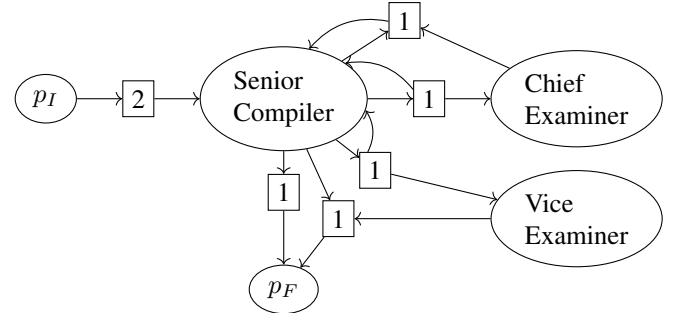


Fig. 3. Model for two sample officials career paths ( $L_E$ ) discovered by the State Snapshot Miner.

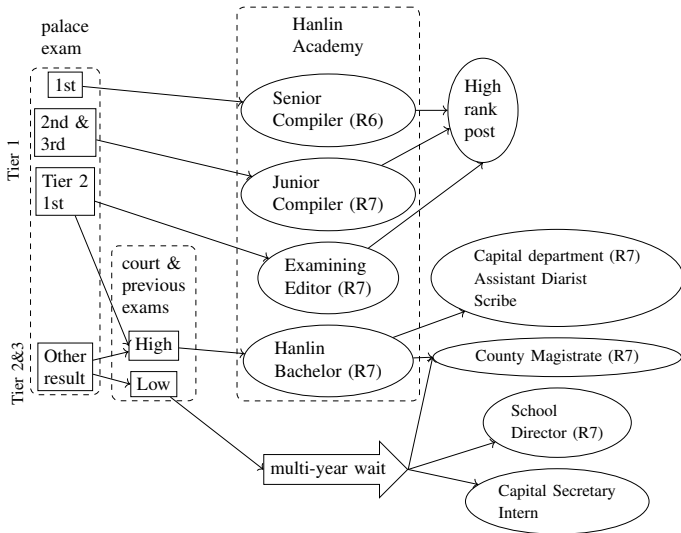


Fig. 4. Flow chart of appointments for *jinshi* 進士 degree holders after taking the palace exam, adapted from [9, p96]. Ranks are shown by (Rn), with R1 being the most senior.

a threshold interpreted as a proportion of the sum of all direct-follow role set pairs (i.e., transition weights).

The visualisation for Place-labelled Petri nets developed for this project uses conventional ovals for places and rectangles for transitions. Arc width is scaled with transition weight, and place size scaled with role frequency. Visualisation support allows the inclusion or suppression of final places.

## V. QING PROMOTION PATHS CASE STUDY

These techniques have been applied to data from the CGED-Q database to examine civil service career paths.

### A. Background

The Qing state had rules for appointments to civil service positions. Notably for our study, many officials were appointed based on performance in civil service exams. There were multiple stages of examination, starting from a prefectural exam, through a provincial exam to national (or Metropolitan) and palace exams. The exams were highly selective, with approximately 2% of the adult male population, or 2 million people, presenting for each prefectural examination, but only 26000 candidates succeeding at the national exam over the entire course of the entire Qing dynasty [2]. Success at each level of these exams led to the award of a degree, with admission to the palace exam granting the highest degree: *jinshi*. Specific rules applied to the appointment of the *jinshi* degree holders, which have been the focus of previous scholarship [2], [3]. These defined an “fast track” of sorts for these elite candidates, where they could be quickly placed in relatively senior roles.

The flow-chart in Figure 4, adapted from [9, p96], summarises the normative process model of this appointment and promotion process. Though the candidates had already succeeded at three exams, a further ranking was provided by a palace exam, hosted by the emperor himself. Candidates were

TABLE II  
CGED-Q LOG SUBSET STATISTICS, 1830-1904

Palace Exam Filter	Year Filter	State Entries	Cases	Max Concurrent Roles
Top place	None	375	36	18
Top place	3	64	36	3
Tier 1 & 2 Top 7	None	11790	3897	5
Tier 1 & 2 Top 7	3	5299	3897	4

ranked into three tiers, with the top three candidates constituting Tier 1 (一甲). These three candidates, and sometimes the top placed candidate in Tier 2, were given posts at the Hanlin Academy and groomed for high office. The Hanlin Academy was an elite academy that served the imperial court. The remaining Tier 2 and 3 candidates sat a further exam, the court exam, which determined admission to the Hanlin Academy, or appointment to less prestigious, but still important, roles in the capital and the provinces. In Figure 4 we can see, for example, that the top candidate in the palace exam was placed into a Senior Compiler post (修撰), part of a team which drafted and compiled official histories. Being the top graduate earned a specific title (状元) and the distinction was noted in the *jinshenlu* and other records.

### B. Data Preparation

To examine elite graduate Qing civil service careers from a process perspective, we used an extract from the CGED-Q database that covered ethnic Han officials who earned the *jinshi* degree and sat the palace exam over the period 1830-1904. The CGED-Q records are most complete after 1830, and the exams were abolished in 1905. This period was an eventful one in China, encompassing the Taiping Rebellion and both Opium Wars. In the current version of CGED-Q, the challenge of uniquely identifying officials over time, using a combination of name and other identifiers, has been most completely solved for ethnic Han officials, in part due to the relative uniqueness of the names of elite males in the period [22]. CGED-Q has established a unique `person_id` field as a result of this work. We joined the *jinshenlu* extract with the examination records from the same period (會試題名錄) to obtain the palace exam tier. Officials who were already mid-career in 1830 were excluded.

Concurrent roles for an official were sometimes represented as separate records in the same circular, and sometimes concatenated in a single field in the *jinshenlu* record. This is an instance of the *Distorted Label* event log imperfection pattern [23], a variant extreme enough to result in role conflation. To resolve this, the role label was split according to a new project dictionary of known roles. Concurrent roles often had partially overlapping time periods and different durations. For example, an official may have a permanent posting, which lasted a number of years, a temporary appointment, lasting a few months, and an honour or title which endured. Honours, such as “gifted the peacock feather” 賞戴花翎, were similar to knighthoods in the British system. A number of synonyms

for roles were also remapped to a common role name, in an instance of *Synonymous Labels* [23]. For example, those allowed to study at the Hanlin Academy (Hanlin bachelors 庶吉士) often had the province name or graduating class name included in their role title, and this was removed. Chronological gaps in sequential appointments, which might indicate unemployment, were not distinguished from other sequences of states in this log. For elite early career officials in this period, unemployment would be rare, but gaps of approximately six months do exist in the CGED-Q database, when the original circular for that season was not available for entry into the database. Lastly, continuing role combinations were conflated into a single record, and this was formatted as a comma separated value (CSV) file that could be parsed as a State Snapshot Log as described in Definition IV.3. The `person_id` was used as a case id, the job title as the role, and the year and month as a timestamp. The resulting log consists of 219276 individual appointment records. Table II lists statistics for key subsets of this log.

### C. Early Career Path Models

A variety of models based on different data selections and periods of appointment were produced for analysis and discussion during the study. Some example insights into the domain gained with the State Snapshot miner can be explained using the models in Figures 5 and 6. The state log was filtered by time to produce logs for the first few years of an official’s career. The State Snapshot Miner was then applied to a variety of filtered subsets for different exam tiers and career segments. All diagrams for this case study exclude the final place, and the transitions leading to it ( $T_F$ ), for clearer visualisation. Arcs and places are also scaled by frequency. Conceptually, given the domain, all career stages are potentially final markings. The undisplayed transitions and weights are available in the backing Place-labelled Petri net model.

Figure 5 is a model of the promotion paths for the top-placed candidate in the palace examination over the first three years of their career. Thirty-six officials are in this category, allowing for fine-grained models. We can see that this confirms the normative model in Figure 4, showing these officials appointed as Senior Compilers within three years, and often earlier. One exceptional official was instead appointed to a privileged position as Household Administrator of the Heir Apparent, still fulfilling the promise of a fast track. There are also elements not shown in the simplified normative model, such as officials serving concurrent roles as provincial interns or examiners. Such officials sometimes served briefly in junior roles before starting their Senior Compiler appointment.

Figure 6 is a model of the 3897 officials achieving Tier 1 or 2 in the palace exam, over the first three years of their career. To obtain a comprehensible model, the seven most popular roles were retained, and the remainder replaced with their civil service rank. Ranks in CGED-Q are identified at a three band granularity, e.g. 4-6, so these conflated roles are shown as other-4-6. Noise reduction of 0.08% was also applied. Figure 4’s normative appointment model reflects

known regulations for appointment, but the degree to which these rules were actually adhered to is not well established. The Figure 6 model largely confirms the normative paths were respected, while showing a number of idiosyncratic variations occurred in practice. A number of Tier 2 officials are admitted to the Hanlin Academy as Hanlin Bachelors. 365 proceed onto Junior Compiler positions, while many others are appointed as Secretaries. Those appointed as County Magistrates or Prefects do not move beyond that within this two-year period. Lastly, more senior (and diverse) rank 4-6 openings were possible directly out of the palace examination without necessarily placing in the first tier.

For these logs, execution of the miner completes in seconds on a Windows 10 machine with a 2.3GHz CPU machine, 10 Gb memory, and Python 3.8.

## VI. DISCUSSION

### A. Domain Expert Insights

Development of the data preparation, discovery algorithm, and visualisation was an iterative and collaborative process in a team including both technologists and Qing historians, some of whom were already familiar with computational data science tools, and some of whom were relative novices. We conducted structured reflections in the team to consolidate our understanding of the impact and potential of these process mining techniques.

In general, the historians understood the meanings of directed graph visualisations quickly. Historians also contributed design suggestions on visual elements that were incorporated in the tool. The (conventional) Petri net visual representation of concurrency, though understandable, was also noted as complex or “cluttered” when temporary and long-term roles were in parallel.

The value of comparison with the normative model in Figure 4 was highlighted by historians, and seen to complement existing domain methods. “There’s a lot of things in eighteenth and nineteenth century China where there’s deviation between the regulations and the actual practice, so actually showing that practice largely followed the [...] regulations [...] is certainly quite important.” The identification and visualisation of typical versus exceptional pathways was seen as a particular strength.

It was noted that Tier 1 officials, such as the subset in Figure 5, have already attracted significant scholarship using qualitative close reading techniques (surveyed in [2]). Models from larger data sets, such as the Tier 1&2 combined model in Figure 6, were reported as new. “When you get into categories of people that are too numerous to subject to case studies, then it really gets interesting to have that [...] zoomed out summary view.” They also see the techniques as having potential for discovering hidden rules for appointments that are not documented in official manuals.

### B. Representation and Limitations

The current miner does not support multiple places having the same role (the equivalent of duplicate labels), or silent places with no observed role. This is suitable for career path

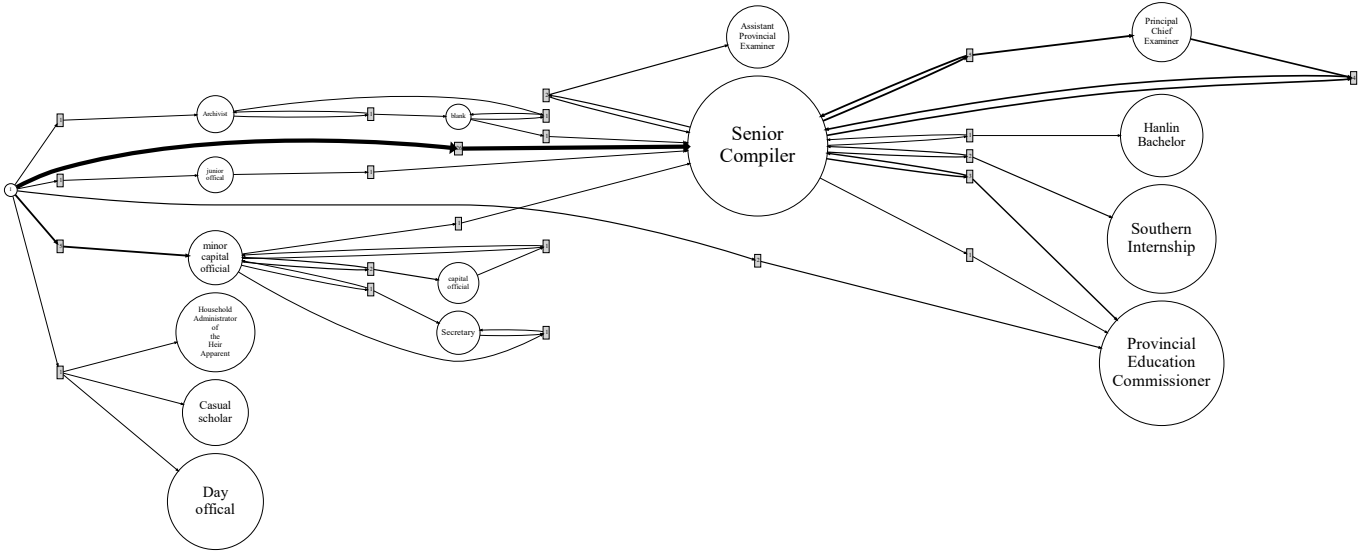


Fig. 5. Roles held by officials who placed first in the palace exam (状元), in the first three years of their career, 1830-1904.

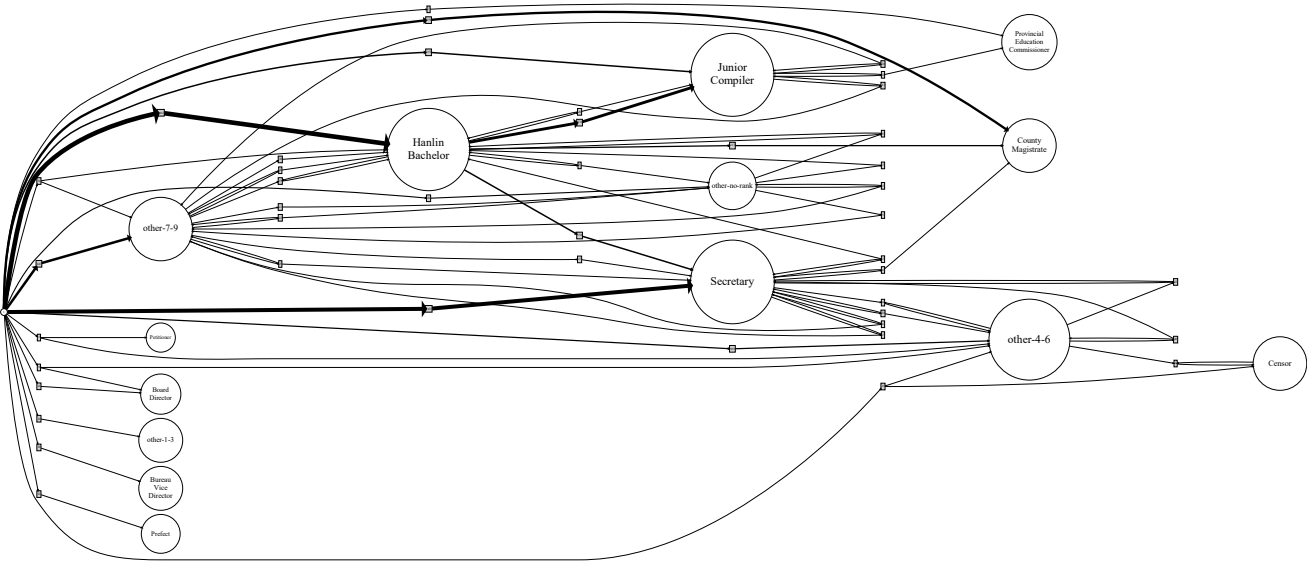


Fig. 6. Roles held by officials placed in Tier 1 or 2 in the palace exam, in the first three years of their career, 1830-1904. Top seven roles with rank conflation on remainder, noise reduction 0.08%.

settings including the CGED-Q data, where such duplicate or silent roles would not be meaningful, but could constrain its use on state-snapshot data from other domains.

The miner gives a way to represent models in a diagram with a well-defined formal semantics derived from stochastic Petri nets. Other representations are possible. For example, a decision tree, or Stochastic Deterministic Finite Automata (SDFA) [24], could also represent all of the possible states for a set of official careers. In the case of a decision tree, representing the variation in paths through particular roles would result in a large number of entities. For an SDFA or Direct Follow Graph, the number of entities would be smaller, but each unique state would have to correspond to multiple official appointments. Though the state space is finite, its size

would still explode as roles and path variations increased. A Petri net, where the state is represented by a marking - tokens in multiple places - allows for a more concise representation of both state and concurrent roles.

A limitation on the accuracy of the algorithm comes from using direct-follows frequencies to provide weights. This can over-represent the probability of such transitions in loops. E.g. the trace  $\langle \{b\}, \{b\}, \{b\} \rangle$  will contribute 2 to the weight of the  $(\{b\}, \{b\})$  transition, and the  $b$  place will occur in a loop, causing a “double counting” effect.

## VII. CONCLUSION

The Qing civil service case study shows that process mining concepts and techniques can be applied well beyond modern



information technology settings. Process mining traces its antecedents to an industrial process improvement tradition, including the scientific management of Frederick Taylor, the assembly line of Henry Ford, and the twentieth century advent of modern computing [1, p55]. The Qing civil service, and its records, are from a different culture that predates all of these social changes. Historical, paper-based bureaucracies are also information technologies, and the behaviour of an organisation may still be understood through its processes, especially if the records are available in a structured digital form.

The setting still presented significant research challenges arising from the structure of the data. To meet these, we have introduced state snapshot logs, a discovery algorithm, the State Snapshot Miner, for discovering stochastic process models from such logs, an implementation, and a visualisation of the resulting models. Models of the early career paths of elite officials show conformance to the bureaucratic rules of appointment, and variation and complications in how such rules were followed in practice. For instance, promised appointments may be awarded after a short period in less prestigious roles, even after an official had achieved exceptional performance in examinations. These insights are difficult to obtain using qualitative close reading methods, which focus on individual officials, or quantitative statistical methods insensitive to paths.

Looking forward, state snapshot mining can also be used in a more exploratory mode, especially if the tools are extended to allow interactive data selection, constraint, and filtering. Other uses of the miner might include Internet of Things (IOT) sensor data, where multiple sensors capture the state of a system at intervals. More broadly, progress by historians applying computational and quantitative methods has made many historical datasets available, and we see understanding this “long history of information systems” as a rich and promising area of research.

#### ACKNOWLEDGEMENTS

We wish to thank Dr Bijia Chen and Prof Arthur ter Hofstede for their expertise and insights, as well as the anonymous reviewers. While participating in this study, Cameron Campbell was a 2022-23 Fellow of the Stanford Center for Advanced Study in the Behavioral Sciences and received support from Hong Kong Research Grants Council General Research Fund 16602621 (Campbell PI) and Areas of Excellence AoE/B-704/22-R (Chen PI).

#### REFERENCES

[1] W. van der Aalst, *Process Mining: Data Science in Action*, 2nd ed. Berlin Heidelberg: Springer-Verlag, 2016.

[2] B. A. Elman, *Civil examinations and meritocracy in late imperial China*. Harvard University Press, 2013.

[3] 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, Jul. 2020.

[4] C. Campbell and J. Lee, “Historical Chinese Microdata. 40 Years of Dataset Construction by the Lee-Campbell Research Group,” *Historical Life Course Studies*, Sep. 2020.

[5] C. Cioffi-Revilla, “Computational social science,” *WIREs Computational Statistics*, vol. 2, no. 3, pp. 259–271, 2010.

[6] A. T. J. Barron, J. Huang, R. L. Spang, and S. DeDeo, “Individuals, institutions, and innovation in the debates of the French Revolution,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 18, pp. 4607–4612, May 2018.

[7] Harvard University, Academia Sinica, and Peking University, “China Biographical Database (CBDB),” Jan. 2023.

[8] C. O. Hucker, *A Dictionary of Official Titles in Imperial China*. Peking University Press, 2008.

[9] B. Chen, “Origins and Career Patterns of the Qing Government Officials (1850-1912): Evidence from the China Government Employee Dataset-Qing (CGED-Q),” PhD Thesis, Hong Kong University of Science and Technology, 2019.

[10] J. Yang, C. Ouyang, W. M. P. van der Aalst, A. H. M. ter Hofstede, and Y. Yu, “OrdinoR: A framework for discovering, evaluating, and analyzing organizational models using event logs,” *Decision Support Systems*, vol. 158, p. 113771, Jul. 2022.

[11] B. Cornwell, *Social sequence analysis: Methods and applications*. Cambridge University Press, 2015, vol. 37.

[12] A. Abbott and A. Hrycak, “Measuring resemblance in sequence data: An optimal matching analysis of musicians’ careers,” *American journal of sociology*, vol. 96, no. 1, pp. 144–185, 1990.

[13] S. J. J. Leemans, W. M. P. van der Aalst, T. Brockhoff, and A. Polyvyanyy, “Stochastic process mining: Earth movers’ stochastic conformance,” *Information Systems*, vol. 102, p. 101724, Feb. 2021.

[14] Y. Wang, H. Liang, X. Shu, J. Wang, K. Xu, Z. Deng, C. D. Campbell, B. Chen, Y. Wu, and H. Qu, “Interactive Visual Exploration of Longitudinal Historical Career Mobility Data,” *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021.

[15] H. Reijers, “What Have the Romans ever Done for Us? The Ancient Antecedents of Business Process Management,” in *Business Process Management*, Rome, 2021, keynote.

[16] N. Silva, M. Silva, C. Mendes, and M. M. da Silva, “A Public Organization Career Progression Analysis Using Process Mining,” in *Atas da Conferência da Associação Portuguesa de Sistemas de Informação*, vol. 15, 2017, pp. 270–290, issue: 15.

[17] E. Bellodi, F. Riguzzi, and E. Lamma, “Probabilistic Declarative Process Mining,” in *Knowledge Science, Engineering and Management*, ser. LNCS, Berlin, 2010, pp. 292–303.

[18] S. J. J. Leemans, “Leveraging frequencies in event data: a pledge for stochastic process mining,” in *Workshop on Algorithms and Theories for the Analysis of Event Data*, 2022.

[19] S. J. J. Leemans, L. L. Mannel, and N. Sidorova, “Significant stochastic dependencies in process models,” *Information Systems*, p. 102223, May 2023.

[20] F. Mannhardt, S. J. Leemans, C. T. Schwanen, and M. de Leoni, “Modelling Data-Aware Stochastic Processes-Discovery and Conformance Checking,” in *Petri Nets*, 2023.

[21] J. Desel and W. Reisig, “Place/transition Petri Nets,” in *Lectures on Petri Nets I: Basic Models: Advances in Petri Nets*, ser. LNCS. Springer, 1998, pp. 122–173.

[22] C. Campbell, B. Chen *et al.*, “Nominative linkage of records of officials in the china government employee dataset-qing (cged-q),” *Historical Life Course Studies*, vol. 12, pp. 233–259, 2022.

[23] S. Suriadi, R. Andrews, A. H. M. ter Hofstede, and M. T. Wynn, “Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs,” *Information Systems*, vol. 64, pp. 132–150, Mar. 2017.

[24] E. Vidal, F. Thollard, C. de la Higuera, F. Casacuberta, and R. Carrasco, “Probabilistic finite-state machines - part I,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 7, pp. 1013–1025, Jul. 2005.