GMR
The Original

KNOWLEDGE GRAPHS IN DIGITAL HUMANITIES: A SEMANTIC FRAMEWORK FOR CULTURAL KNOWLEDGE DISCOVERY

QiuyingLi[1]

School of English and Advanced Translation, Beijing Language and CultureUniversity,Beijing, 100080,China,louiselee@pku.org.cn

ABSTRACT

Purpose: This study proposes a semantic knowledge graph framework for cultural knowledge discovery in digital humanities using the China Biographical Database (CBDB). The framework supports semantic organization, relationship interpretation, and evidence-based knowledge discovery from large-scale historical records.

Methodology/Approach: A design science and secondary-data methodology was adopted. The framework integrates relational profiling, semantic modeling, spatial concentration analysis, and multiplex network analysis to construct and evaluate a semantic knowledge graph. Nine summaries of the complete database (659,593 persons) and five analytical subsets (5,000 persons) were examined using degree, strength, PageRank, approximate betweenness, Louvain community detection, Spearman correlations, and sensitivity analysis.

Findings: The semantic knowledge graph demonstrates broad relational coverage while revealing temporal incompleteness and uneven information distribution. The analytical graph contains 18,810 semantic relationships and 359 knowledge communities. PageRank shows moderate consistency with heterogeneous record prominence ($\rho = .465$), with stable rankings across weighted specifications but noticeable variation within the kinship layer. These findings demonstrate the effectiveness of semantic network analysis for uncovering implicit cultural knowledge and structural patterns in historical datasets.

Originality/Relevance: The proposed framework extends knowledge graph applications in digital humanities from visualization to semantic knowledge discovery. By emphasizing semantic organization, relationship analysis, and methodological transparency, it provides a reproducible approach for cultural analytics, historical network research, and digital humanities scholarship.

KEYWORDS: Knowledge Graphs; Digital Humanities; Semantic Framework; Cultural Knowledge Discovery; China Biographical Database; Historical Networks.

1. INTRODUCTION

The cultural collections are no longer limited to physical access. What is the problem with analytic fragmentation? Names are linked to offices, texts to authors, places to administrative changes and social relations to the documents which bear witness to them. One solution to this is knowledge graphs, which are a representation of typed relationships and entities in a machine-readable format. They are not, however, worked up when published or searched. A graph is analytically consequential if it can be used for clear prioritization, comparison, detecting anomalies and iterative improvement of scholarship questions. The current research on knowledge graphs highlights both representation, acquisition and application, and highlights the need for explicit decisions on scope, semantics and evidence quality within domain-specific knowledge graphs (Abu-Salih, 2021; Hogan et al., 2021; Ji et al., 2022).

The advances of digital-humanities projects in the areas of linked data, semantic portals, and exploratory interfaces have been significant. The Sampo model, for instance, shows how the diverse cultural resources can be unified and presented in a multifaceted manner with search and analytical services (Hyvönen, 2020, 2023; Ikkala et al., 2022). There is related research on metadata for manuscripts, historical urban knowledge, and the work of cultural storytelling, which demonstrates possibilities for semantic integration that could link objects, people, places, and narratives between institutional borders (Koho et al., 2022; Krabina, 2023; Renzi et al., 2023). However, a common challenge continues to be the disconnect between knowledge graph creation and a proven evidence-to-insight workflow. In many studies, an ontology, interface or case application is described but little attention is given to the changes in the resulting cultural interpretation when the coverage is insufficient, records are multiple, subgraphs are selected and alternative layers of relationships are provided.

Competitive intelligence provides a valuable yet potentially deceptive vocabulary to close this gap. Competitive Intelligence in Management Research is the process of systematically scanning the environment, synthesizing information, and converting it into insight that is relevant to decision making. It has been found to be correlated with strategy formulation, organizational analytics, innovation and implementation, but it also has conceptual inconsistencies and poor theoretical integration (Bao, 2020; Cavallo et al., 2021; Maluleka & Chummun, 2023; Maungwa & Laughton, 2023; Ranjan & Foropon, 2021). The term may be transferred without any critical

understanding and therefore convey a sense of competition among cultures or of historical ranking. In this paper, we will take a more scholarly definition: competitive intelligence is a transparent cycle of sourcing, validation, semantic integration, analytical prioritization, interpretation and feedback. Competition is about alternative explanations and alternative specifications of the graph; it's not about market valuations of the cultural actors.

The empirical setting is the China Biographical Database (CBDB), a massive prosopography database for premodern China. CBDB connects a person's name to circumstances like other names, offices, social associations, kinship, places, texts, institutions and events, especially suitable for graph-based inquiry (Chen & Wang, 2022; Tsui & Wang, 2020). Its scope has been a catalyst for networks and spatial analyses, and its structure has made it possible to analyze things that would be hard to do with individual biographies (Bol, 2020; De Weerdt, 2020; Fuller & Wang, 2021). Meanwhile, the database is a growing academic infrastructure, and not a census. There is some variation in record density across dynasties, between the various relationship domains, among the different source traditions and in the coding history. Documenting a count as a population frequency would be confusing, since visibility in the past does not indicate frequency of occurrence.

To meet that demand, this study proposes a knowledge graph-based cultural knowledge discovery system based on a competitive intelligent algorithm. It brings together five elements: release and schema validation; full-database coverage profiling; building a heterogeneous person-centred knowledge graph; multiplex social and kinship network analysis; and robustness testing under varying weighting of the edges of the graph, varying compositions of the layers of the graph and varying treatment of the outliers. The framework also defines a transparent knowledge-prominence score using the concept of knowledge visibility through the use of heterogeneous visibility. This score is not displayed as a latent entity, as a marker of human significance and/or as a causal result. It is a repeatable tool to choose an information-rich subgraph of the analytical graph and to compare how the visibility of the graph in archives correlates to its prominence in the graph structure. The result is the evidence-to-insight architecture and governance controls that span each phase of the process (summarized in Figure 1).

Four questions are asked in the study. First, is CBDB coverage wide and balanced in relational/ temporal aspects? Second, who are some of the historical figures that have structurally strong positions in the selected multiplex network and what is their position relative to heterogeneous record visibility? Third, are social and kinship layers ranked according to the same cultural-intelligence? Fourth, are the ranks consistent across different weighting and outlier specifications? The answers to these questions are offered in this paper as much as anything else as a method the writer can use again and again as a tool for understanding the Chinese past, not as a cause and effect history. The findings delineate the recorded relational structure, surface high connectivity cultural actors and layers of evidence worthy of deeper humanistic interpretation.

The impact is four-fold. Theoretically, the paper brings the competitive-intelligence process logic together with the knowledge discovery of digital-humanities and resists the incorrect application of market-performance assumptions. Empirically, it analyzes 659,593 person records of CBDB and several relation domains that are linked to each other, in a reproducible pipeline. The methodologically distinguishes full database profiling from a bounded high information network; quantifies the convergent evidence between topology and record visibility; and performs a layer specific sensitivity analysis on rankings. From the perspective of Chinese prosopography data, it introduces the explicit treatment of the aspects of dynasty, kinship, office, and place and source coverage asymmetry. This framework is designed to supplement, not supplant, close reading and domain knowledge.

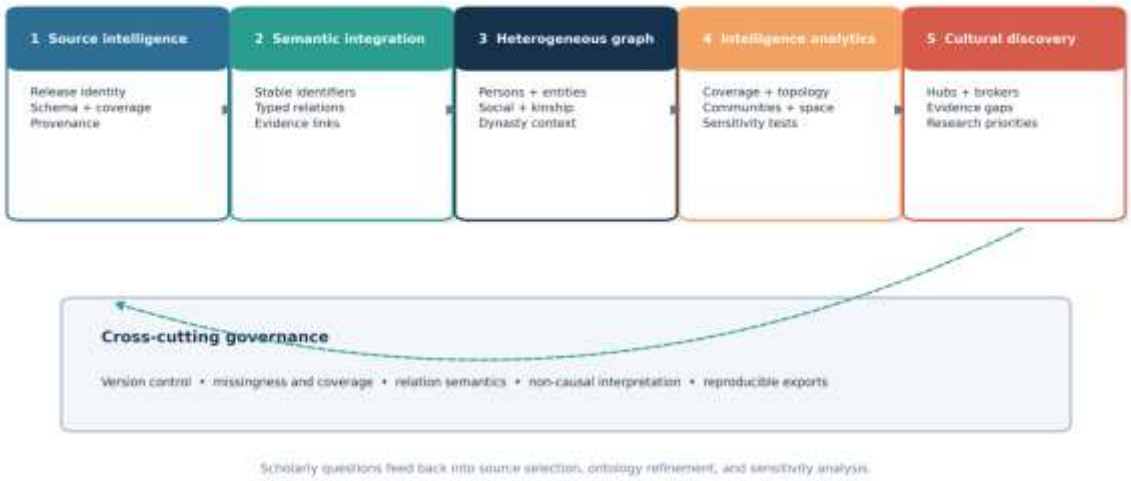

Figure 1. Competitive intelligence-driven framework for cultural knowledge discovery.
Source: Developed by the authors from the analytical framework.

2. THEORETICAL FRAMEWORK

2.1 COMPETITIVE INTELLIGENCE AS AN EVIDENCE-TO-INSIGHT PROCESS

Competitive intelligence tends to fall under the umbrella of organizational strategy. It involves several core

activities: articulating information needs; observing an environment; verifying sources; synthesizing divergent data; creating interpretation; and putting the interpretation into decisions. The recent literature is not as consistent in terms of theory as it is in terms of process. Cavallo et al. (2021) demonstrate that intelligence is useful when it is used as part of a strategy formulation process and not simply as an information service. Similarly, Ranjan and Foropon (2021) highlight the need for large-scale analytics to generate value when data are transformed into an organization's knowledge that is interpretable. While Bao (2020) states that the concept of competitive intelligence has been found to be empirically attractive in the context of innovation in the Chinese tourism industry, he also notes its traditional association with the market outcomes.

2.2 KNOWLEDGE GRAPHS IN DIGITAL HUMANITIES

A knowledge graph is a graph that models entities as nodes and semantically typed relationships as edges. It's a seemingly simple structure that can handle a variety of tasks, such as entity exploration, answering questions, predicting links, reasoning, etc. There are general surveys that separate the tasks of graph construction, representation learning, completion, and downstream tasks, and domain-specific studies that focus on ontology design, as well as aligning local concepts with external vocabularies (Abu-Salih, 2021; Hogan et al., 2021; Ji et al., 2022). The key to the cultural research is to hold on to the evidence heterogeneity. They can relate a person to a place, office, text, institution, event, kin relation, and social associate at the same time and not collapse all of the relations into one rectangular table.

Other studies demonstrate the diversity of cultural-heritage uses. O'Neill and Stapleton (2022) provide reasons for shifting standards from digitization to semantic interoperability. Casillo et al. (2023) present an ontology-based platform to share cultural knowledge and Renzi et al. (2023) merge linked open data, knowledge graph, and neural approaches to multimedia storytelling. In Krabina (2023), the historical knowledge of a city is shown to be an evolving graph instead of a single story. In a recent article in the International Journal of Human-Robot Interaction, Chen et al. (2023) demonstrate the utility of character-network maps to enable digital-humanities inquiry by making social structure inspectable. These can all be seen as useful, but do not preclude the need to assess analytical stability and archival bias. In the context of China-related heritage research, the multimodal intangible-heritage graph, the bilingual martial-arts graph and the ontology-based embroidery graph are examples of the growing number of domain-specific cultural models in China-related research (Fan et al., 2023; Hou & Yuan, 2023; Liang et al., 2025).

Evaluation is further reinforced by the research in the field of knowledge-graph embedding and completion. Rossi et al. (2021) prove that model classes and datasets can differ in their link-prediction performance, while Ali et al. (2022) find that the implementation options, losses, inverse relationships, and training configurations can significantly impact the benchmark results. Event knowledge graph is another challenging aspect due to the distinction between temporal and causal semantics from co-occurrence (Guan et al., 2023). In the present study, no link prediction is performed - only taking the lesson from the evaluation lesson that a graph result should not be trusted without specification checks. This involves making weightings and unweighting projections for historical networks, differentiating between kinship and social association, and establishing whether extreme record multiplicities outweigh rank order in historical networks.

2.3 CHINESE HISTORICAL DATA AND CULTURAL KNOWLEDGE GRAPHS

CBDB is geared towards relational historical research. Its usefulness as a reference work is not just for the number of individuals it contains but for the clear indication of many-to-many relationships. The database has subsequently been introduced as an infrastructure for harvesting large-scale biographical data by Tsui and Wang (2020) and as a relational database and for research purposes by Chen and Wang (2022). Historical networks can be structured, recorded and analyzed in CBDB, in the process revealing that a network edge is not a physical location but rather a product of interpretation by sources and coding in the database. (Fuller and Wang 2021: 107–108). De Weerdt (2020) places CBDB into context of tools for the production and connection of historical datasets of Chinese origin.

When these mechanisms of construction and selection of large historical databases are respected, they can be used in substantive historical research. Wen et al. (2024) use large-scale historical evidence to examine social mobility during the Tang dynasty, and make clear links between analytical measures and historically relevant institutions. A lesson in knowledge-graph research is that there will always be a need for historical interpretation, regardless of the scale. Being a node of a network can mean that a person has a large number of recorded relations with others, but it does not automatically imply ideological influence, social status or causal power. Those interpretations need to be done through source reading, period expertise and triangulation.

2.4 RESEARCH GAPS

This is where the gaps come together around a central problem: there's a need for an analytical governance layer for cultural knowledge discovery. A graph can be technically correct but still give false impressions historically because of missingness, selection and relation semantics are hidden. In accordance with this, the data quality and sensitivity analysis is viewed as an integral part of the substantive result in the framework set out here. A low coverage relation domain is not just a mere technical hindrance; it is a place where the archive does not provide for confident discovery very well.

Following this problem, there are four related problems. Conceptually, research into competitive intelligence is predominantly located within an organizational and market context, while cultural knowledge infrastructures are

based on a non-commercial perspective of how evidence is made into a defensible research priority (Cavallo et al., 2021; Maluleka & Chummun, 2023; Maungwa & Laughton, 2023). In practice, cultural knowledge-graph studies have not been shown to reproduce gender and source asymmetries that can be seen in the archive. Methodologically they often also report the composite visibility scores and multiplex projections without trying alternative weights of the domains or layers of relationships. From a contextual perspective, photographical information from China is compounded by generally shifting place authorities, dynasty-specific offices, kinship semantics, and the lack of sources for certain individuals (Tsui & Wang, 2020; Fuller & Wang, 2021). The contribution is therefore limited to a narrow view of transparency of priorities, provenance-aware interpretation, and sensitivity testing, instead of the concept of a universal theory of cultural intelligence.

2.5 ANALYTICAL PROPOSITIONS

These are analytical propositions rather than causal hypotheses. They specify observable relationships among graph metrics and database-derived indicators. No treatment, counterfactual, exogenous shock, or measurement model is available; accordingly, the study does not estimate effects, mediation, moderation, or causal mechanisms.

P1. Structural network prominence will be positively but imperfectly associated with heterogeneous archival visibility; PageRank and KPS should converge without becoming interchangeable.

P2. Social-association and kinship layers will yield substantively different prominence rankings because they encode different forms of historical proximity.

P3. The combined-network ranking will remain stable when record-multiplicity weights are removed and extreme weights are capped, provided that it is not an artifact of coding multiplicity.

P4. Dynastic and spatial concentrations will be observable in the recorded knowledge structure, but their interpretation will remain conditional on database coverage, missingness, and source selection.

3. METHODOLOGY

3.1 Study Design

The study is carried out based on a design-science and computational digital-humanities approach. The artefact is a repeatable framework which transforms a relational historical database into a governed cultural knowledge graph and a set of interpretable analytical results. Evaluation takes place at four levels: Data Integrity, Descriptive Coverage, Network Structure and Robustness. However, the large dataset has not been exploited in a predictive model, because it is not done automatically. Research questions involve representation, visibility, and relational organisation; descriptive statistics of the graphs, and sensitivity analyses are more suitable than regression or SEM.

There are two analytical branches in the workflow. The first performs all schema validation, missingness, relation coverage, dynastic composition, gender coding, place concentration and relation-domain summaries based on the complete database of 659,593 persons. The second employs a subgraph of 5,000 individuals with a high information level for more computationally demanding centrality and community analyses. This isolation allows the chosen network not to be confused with the entire database. It also clarifies that certain results are population of record summaries and others are based on a prioritization rule. The process of validation, profiling, graph construction, and splitting of two analytical branches is depicted in Figure 2.

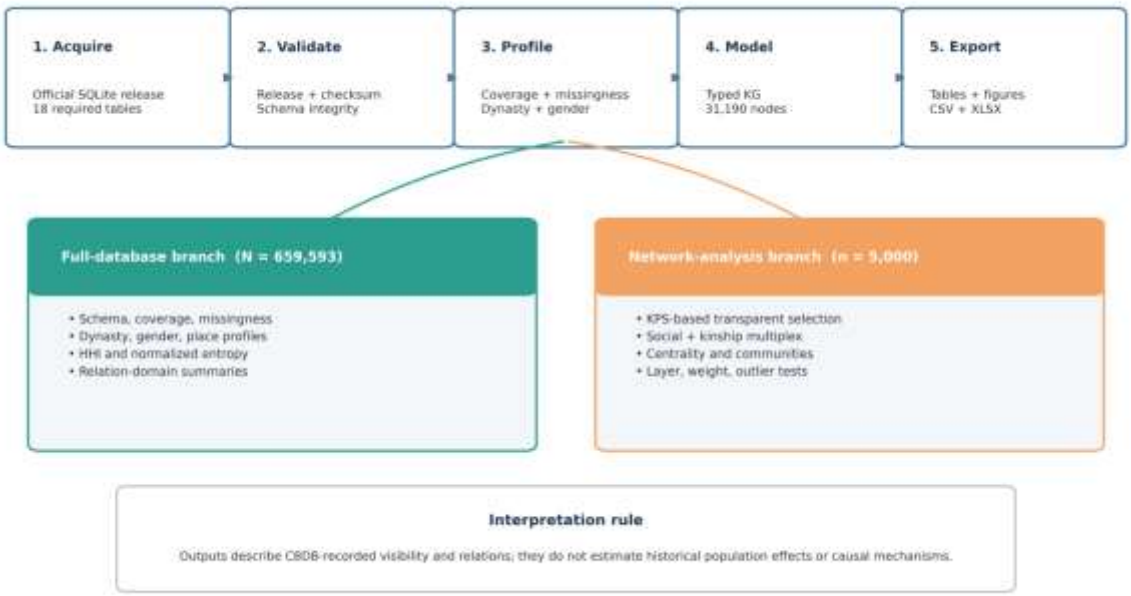

Figure 2. Methodological workflow and separation of the full-database and network-analysis branches. Source: Authors' analysis of the CBDB outputs.

3.2 DATA SOURCE, RELEASE VALIDATION, AND SCOPE

The source is the official CBDB SQLite distribution resolved at analysis time. The runtime database was named cbdb_20260627.sqlite3 and occupied 576,692,224 bytes. The accompanying metadata file listed cbdb_20260314.sqlite3 and a different SHA-256 value, producing a checksum mismatch warning. The analysis therefore records both the metadata-declared release and the actual resolved file rather than silently treating them as identical. This discrepancy does not invalidate the extracted results, but it is a reproducibility issue that should be corrected in the upstream release process or explicitly pinned in a public repository before journal submission.
The unit of analysis changes by task. Person-level summaries use the CBDB person identifier. Network analysis uses person-person dyads derived from social-association and kinship records. The heterogeneous graph includes Person, Place, Office, Text, Institution, and EventType nodes. Spatial concentration uses dynasty-place distributions among persons with a non-missing index place. The scope is the recorded CBDB universe, dominated by Tang through Qing materials. Because the database is curated and historically uneven, results are not generalized to all people who lived in those periods.

3.3 DATA PREPARATION AND SEMANTIC MODEL

Names were not used to determine entity identifiers, but instead, they were carried from CBDB. This option helps lower the false merges of persons with the same name, and keeps connections with source tables. Labels were taken from the available Chinese and romanized name fields. Identifiers were not dropped due to the absence of labels. For the network projection, duplicate relationship rows were aggregated at the dyad level and the resulting edge weight corresponds to the sum of the multiplicity of the relationship rows in the database.
The person's class is the central entity class in the heterogeneous property graph. The following are preserved as typed cultural relations: person-place, person-office, person-text, person-institution and person-event. Person-person relationships are social and kinship relations. Under domain-specific export caps, a portable graph export had 31,190 nodes and 133,383 edges, and analytical aggregates were calculated from the full set of the appropriate tables. The export caps are not a sampling rule for the descriptive analysis, but rather a choice of storage and portability.
For network analysis, directed SRs were transformed to an undirected graph format, as the main goal was not to measure direction of exchanges but to measure relational prominence. The source and target identifiers were ordered in each dyad, a self-loop was deleted and duplicate rows were added together. This projection adds more of an ability to compare the social and kinship layers but loses direction and role semantics. Edge weights are not necessarily the number of times that interactions, affection, influence or communications have occurred in the past.

3.4 MEASURES AND OPERATIONALIZATION

Table 1. Variable definitions and operationalization

Construct	Role	Unit	Operationalization	Scale/interpretation
Knowledge prominence score	Outcome / discovery index	Person	Mean positive percentile rank across seven relation domains, multiplied by 100	0–100 descriptive score; equal-weight baseline with alternative weighting tests
PageRank	Primary network outcome	Person	Weighted PageRank on the undirected social–kinship multiplex graph	Higher values indicate structural prominence in selected graph
Network degree	Network outcome	Person	Number of unique neighbors in the analytical graph	Count
Weighted degree strength	Network outcome	Person	Sum of database-record multiplicities across incident edges	Record-weighted count; not interaction intensity
Approximate betweenness	Network outcome	Person	Approximate betweenness, k = 200, seed = 42	0–1 normalized approximation
Community membership	Network outcome	Person	Louvain partition on the weighted analytical graph	Nominal community identifier
Spatial HHI	Contextual outcome	Dynasty -place distribut ion	Sum of squared place shares within each dynasty	0–1; higher means greater concentration

Construct	Role	Unit	Operationalization	Scale/interpretation
Normalized spatial entropy	Contextual outcome	Dynasty -place distribut ion	Shannon entropy divided by log of the number of observed places	0–1; higher means greater dispersion
Relational activity indicators	Input indicators	Person	Counts of association partners, kin partners, office, text, institution, event, and place records	Non-negative counts
Dynasty, century, gender, index place	Contextual strata	Person	CBDB-coded historical attributes	Used for stratification and interpretation, not causal controls

Note: The measures are archival counts, graph statistics, or contextual indices; psychometric reliability and validity coefficients are therefore not applicable.

Table 1 defines the study measures and their interpretation. The knowledge-prominence score (KPS) provides a transparent person-level summary of heterogeneous database visibility. Seven positive count indicators were used: unique social partners, unique kin partners, office records, text records, institution records, event records, and place records. Within each indicator, positive values were converted to percentile ranks; structural zeros were excluded from the within-domain percentile calculation and retained as zero. For person i, J_i denotes the set of seven domains and PR+_ij the zero-adjusted positive percentile rank in domain j:

$KPS_i = 100 \times (1 / |J_i|) \times \Sigma_{j \in J_i} PR^{+}_{ij}$

KPS ranges from 0 to 100. Equal weighting was chosen to avoid claiming that one archival domain is intrinsically more important than another. The score is descriptive and prioritizing, not psychometric. A high score indicates broad or intense representation across available CBDB relation domains; it does not mean that a person was historically superior, more influential, or more culturally valuable.

Equal weighting is itself an analytical choice. A bounded sensitivity test was therefore performed on the exported top-1,000 person-metric table. Three theory-driven alternatives were compared with the equal-weight baseline: a coverage-focused score assigning 0.20 to association, kinship, office, text, and place domains while excluding the extremely sparse institution and event domains; a network-oriented score assigning weights of 0.30, 0.25, 0.15, 0.10, 0.05, 0.05, and 0.10 to association, kinship, office, text, institution, event, and place; and a documentation-oriented score assigning 0.10, 0.10, 0.10, 0.25, 0.15, 0.15, and 0.15. Spearman correlations quantify local rank stability. Because component-level metrics were exported for 1,000 persons rather than the full database, this test is reported as a bounded diagnostic and not as a complete re-ranking of all 659,593 records.

Network degree counts unique neighbors. Weighted degree strength sums incident edge weights. PageRank measures recursive prominence: a node receives more weight when it is linked to other prominent nodes. For a weighted undirected graph G and a damping parameter d, the score is represented conceptually as:

$PR(i) = (1 - d)/n + d \times \Sigma_{j \in N(i)} [w_{ji} / \Sigma_k w_{jk}] PR(j)$

Approximate weighted betweenness estimates the proportion of shortest paths passing through a node using 200 sampled source nodes and edge distance equal to the reciprocal of relationship weight. Louvain community detection was applied to the weighted network with resolution 1.0 and seed 42. Degree, strength, PageRank, and betweenness capture different structural properties; they are expected to correlate but should not be treated as interchangeable.

3.5 DESCRIPTIVE, SPATIAL, AND COVERAGE ANALYSIS

Coverage was measured in percent of the number of individuals having at least one record in a relation domain. The proportion of missing data at the field level was determined for the name, gender, dynasty, index year, index place, birth year, death year, and death age fields. These measures are reported prior to substantive network results because they specify the extent of the support that the database can provide. In particular, it depends on the temporal missingness whether event-history or panel models are feasible.

Gender coding was not analysed as a demographic variable, but as an AR diagnostic. The entire database gender distribution was compared to the gender of the graph selected with deterministic KPS, and median KPS and PageRank were summarized by coded gender. A population hypothesis test was not applied as CBDB is not random and the selection of the analytical graph was based on observation of the visibility of the record.

$HHI(d) = \Sigma_p s(p,d)^2$

Normalized Shannon entropy provides a complementary measure of dispersion:

$H^*(d) = -\Sigma_p s(p,d) \ln[s(p,d)] / \ln[P(d)]$

3.6 NETWORK SELECTION AND GRAPH ANALYSIS

The 5,000 people with the top 5,000 KPS were chosen for the network branch. The aim of selection was computational and substantive – it aims to produce a graph that is dense enough to be amenable to structural analysis, and yet still includes persons with evidence distributed over multiple domains. The subgraph selected

is not random and does not represent all CBDB persons. The advantages of its conclusions are that it tackles a question with a condition: among the information-rich people in this database, who have a prominent place in the documented social and kinship structure?
Selected individuals were socially assisted and kinship edges were introduced and added to an undirected weighted multiplex projection. The basic graph consists of 5,000 nodes and 18,810 different links. The network density, connected components, clustering coefficient, transitivity, assortativity, centrality and community structure were calculated. The disconnected nodes will not be able to take part in the same paths and so the largest connected component was reported separately.
The convergent evidence was evaluated with the help of Spearman rank correlations between degree, weighted degree, PageRank, approximate betweenness and KPS. The correlation is Spearman since the distribution is highly skewed and rank order is more important than linear distance. The PageRank-KPS correlation was estimated using 3,000-person level resamples and a 95% bootstrap confidence interval was calculated. The bootstrap measures sampling variability in the analytical sample (but not the missingness of the databases).

3.7 ROBUSTNESS AND SENSITIVITY ANALYSIS

Four different graph specifications were compared. Record-multiplicity weights were used for the baseline combined social-plus-kinship graph. When a combined graph was used, but the weights were unweighted, it was tested for the rank order. Layer dependence was tested by a social-only graph and a kinship-only graph. The graph was minorized to the 99th percentile to assess the sensitivity of extreme weights. The Spearman correlation and overlap of top-10, top-20 and top-50 baseline PageRank and each alternative were compared.
The second threat addressed by the KPS weighting test is that of conceptual dependency of evidence domain values. Robustness of a graph specification and sensitivity of a KPS to weight are thus considered independently.

3.8 REPRODUCIBILITY AND RESEARCH INTEGRITY

The analysis was executed from the official CBDB SQLite distribution resolved at run time. The actual database file was cbdb_20260627.sqlite3 (576,692,224 bytes), and its archive SHA-256 was 12eaafb47f28bcc50aa1dea8f5ad93e035e7dfe588f5fd13842d7fdf2597bdec. The accompanying metadata described cbdb_20260314.sqlite3 and a different checksum; both identities are retained so that release drift is visible rather than silently ignored. The supplied Colab notebook and standalone Python script generate the Excel workbook, CSV tables, property-graph exports, centrality files, community assignments, and figures.
Deterministic settings were fixed wherever stochastic procedures were used: seed 42 for Louvain detection and approximate betweenness, k = 200 sampled sources for betweenness, resolution 1.0 for Louvain communities, and 3,000 resamples for the PageRank-KPS bootstrap interval. No direct data were collected from living participants. A persistent repository should contain the exact SQLite file, code, dependency specification, and machine-readable outputs before submission; until that archive is assigned, the filename, checksum, and analytical settings above provide the reproducibility anchor (Jacobsen et al., 2020).

4. RESULTS and DISCUSSION

4.1 DATABASE COMPOSITION, COVERAGE, AND MISSINGNESS

The analyzed database contains 659,593 persons. The gender field is complete at the coding level: 576,712 records are male-coded (87.65%), 57,679 female-coded (8.77%), and 23,586 unknown (3.58%). Dynasty coding is nearly complete, with 0.245% missing. The database is concentrated in several large dynastic groups: Qing contains 236,474 persons, Ming 224,807, Song 83,353, Tang 57,476, and Yuan 25,310. These totals should be interpreted as database composition. They reflect the survival, selection, ingestion, and coding of sources as well as historical population.
Relation-domain coverage is broad but uneven. Place records are available for 59.73% of persons, office records for 45.26%, and kinship records for 43.29%. Social associations cover 6.75%, texts 2.71%, institutions 0.0716%, and events 0.0603%. The heterogeneous graph can technically connect all these domains, but the sparsest relations cannot support population-level comparisons without strong qualification. Their value is primarily case discovery and enrichment.
Temporal fields are the principal constraint. Death age is missing for 90.40% of persons, birth year for 86.88%, death year for 85.13%, and index year for 53.37%. The index place is missing for 40.68%. These rates rule out a complete-case longitudinal model without reducing the data to a highly selected minority. They also explain why the present study uses dynasty-level stratification and recorded index-place distributions rather than event-history analysis or person-year panel regression.
The gender diagnostic reveals an additional selection constraint. Female-coded persons account for 8.77% of the full database but only 34 of the 5,000 selected nodes (0.68%), giving a selected-to-full representation ratio of 0.078. Within the selected graph, the median KPS is 51.31 for female-coded persons and 53.91 for male-coded persons; median PageRank is 0.000088 and 0.000106, respectively. These values do not estimate historical gender inequality. They show that a prioritization rule based on cross-domain record visibility sharply attenuates an already uneven archival representation. Figure 3C visualizes the difference, and Table 2 reports the corresponding descriptive indicators.

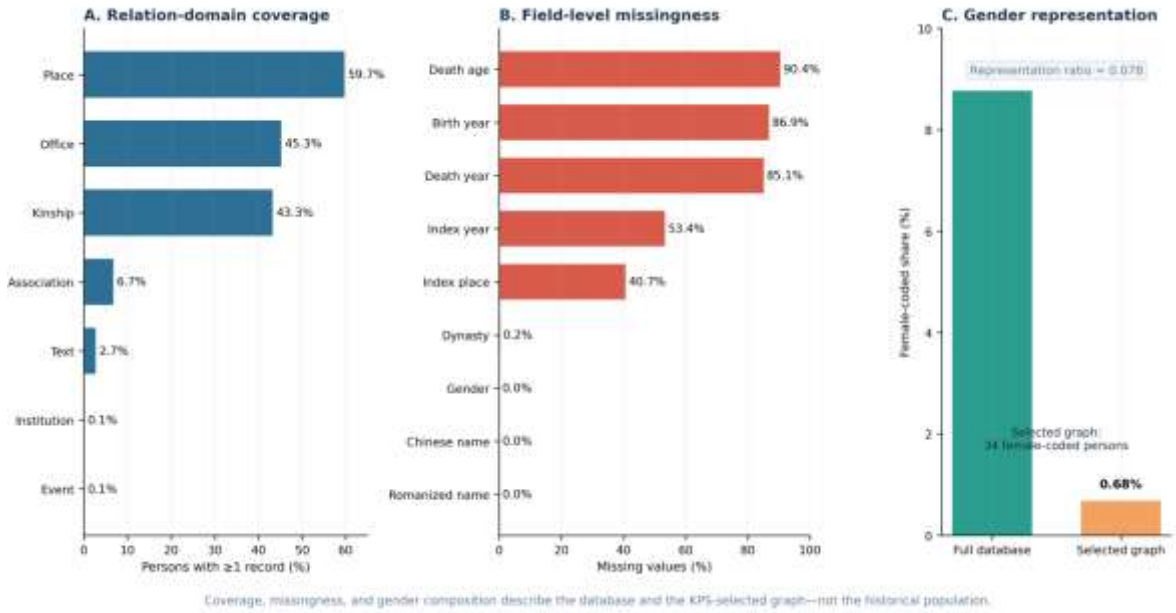

Figure 3. Relation-domain coverage, field-level missingness, and gender representation in the analyzed CBDB release and selected graph.
Source: Authors' analysis of the CBDB outputs.

Table 2. Descriptive statistics and graph summary

Metric	Value	Unit
Persons in full database	659,593	count
Female-coded persons	57,679	count
Male-coded persons	576,712	count
Unknown gender	23,586	count
Female-coded share, full database	8.77	percent
Female-coded share, selected graph	0.68	percent
Selected/full female representation ratio	0.078	ratio
Persons with index place	393,966	count
Persons with office records	298,539	count
Persons with kinship records	285,519	count
Persons with association records	44,493	count
Analytical graph nodes	5,000	count
Analytical graph edges	18,810	count
Graph density	0.001505	proportion
Largest connected component	4,618	nodes
Louvain communities	359	count

Note: Full-database counts use N = 659,593. Network statistics use the selected n = 5,000 graph.

4.2 DYNASTIC AND SPATIAL PROFILES

Spatial concentration varies among the five largest dynastic groups. Qing has HHI = .00336 and normalized entropy = .8489; Ming has HHI = .00401 and entropy = .8231; Song has HHI = .00690 and entropy = .7862; Tang has HHI = .01293 and entropy = .7663; and Yuan has HHI = .00394 and entropy = .8671. On these database-coded index places, Tang is the most concentrated of the five and Yuan the most dispersed by normalized entropy. The leading coded places include Daxing (4,606 persons), Shanyin (2,827), She Xian (2,603), Wanping (2,275), and Putian (2,226). These are archival and coding concentrations; they should not be interpreted as demographic estimates.

The selected graph is also concentrated by dynasty: Ming contributes 1,878 nodes (37.56%), Song 1,808

(36.16%), Qing 672 (13.44%), Tang 350 (7.00%), and Yuan 207 (4.14%). Median PageRank is 0.000098 for Ming, 0.000105 for Song, 0.000108 for Qing, 0.000135 for Tang, and 0.000136 for Yuan. The higher medians for the smaller Tang and Yuan strata are conditional on KPS selection and cannot be read as cross-dynastic influence estimates; they instead indicate that the information-rich cohort has different within-stratum network densities.

4.3 NETWORK TOPOLOGY AND COMMUNITY STRUCTURE

The selected multiplex graph contains 5,000 nodes and 18,810 edges, yielding a density of .001505. Mean degree is 7.524, median degree is 4, and maximum degree is 290. There are 319 isolates and 344 connected components, but the largest component contains 4,618 persons, or 92.36% of the selected nodes. The graph is therefore globally sparse while maintaining a large connected backbone.

Average unweighted clustering is .2106 and transitivity is .0844. The difference indicates substantial local closure around some nodes but a lower global probability that connected triples are closed. Weighted clustering is only .00721 because large record multiplicities are not distributed evenly across triangles. Degree assortativity is negative (−.1063), suggesting that high-degree persons tend to connect with lower-degree persons more often than would occur in an assortative hub-to-hub network. Dynasty assortativity is extremely high (.9004). This value should be interpreted as a combination of historical temporal structure, database coding, and selection: people generally form recorded relations within overlapping periods, and dynasty categories capture much of that temporal separation.

Louvain detection returns 359 communities, including many very small components. In the largest high-information backbone, prominent communities correspond to dense clusters of recorded association and kinship. The visualization in Figure 4 is intentionally restricted to the largest component among the 220 highest PageRank persons; it is a legibility device, not a different analytical sample. Community labels are algorithmic partitions and require historical interpretation before being named as schools, factions, lineages, or intellectual movements.

The largest partitions are strongly aligned with dynastic context. Community 5 contains 899 persons and is 98.8% Ming, led in PageRank by Wang Shouren and Li Dongyang; community 308 contains 875 persons and is 97.7% Song, led by Zhou Bida and Su Shi; community 15 contains 871 persons and is 97.6% Song, led by Zhu Xi and Wei Liaoweng. Communities 3, 10, and 1 are predominantly Ming (96.0%), Qing (91.0%), and Tang (89.0%), respectively. This pattern clarifies the very high dynasty assortativity: the partitions are primarily temporal-dynastic structures before they can be interpreted as schools, factions, or lineages. Figure 4 combines the visible network backbone with the size and dynastic composition of the largest communities.

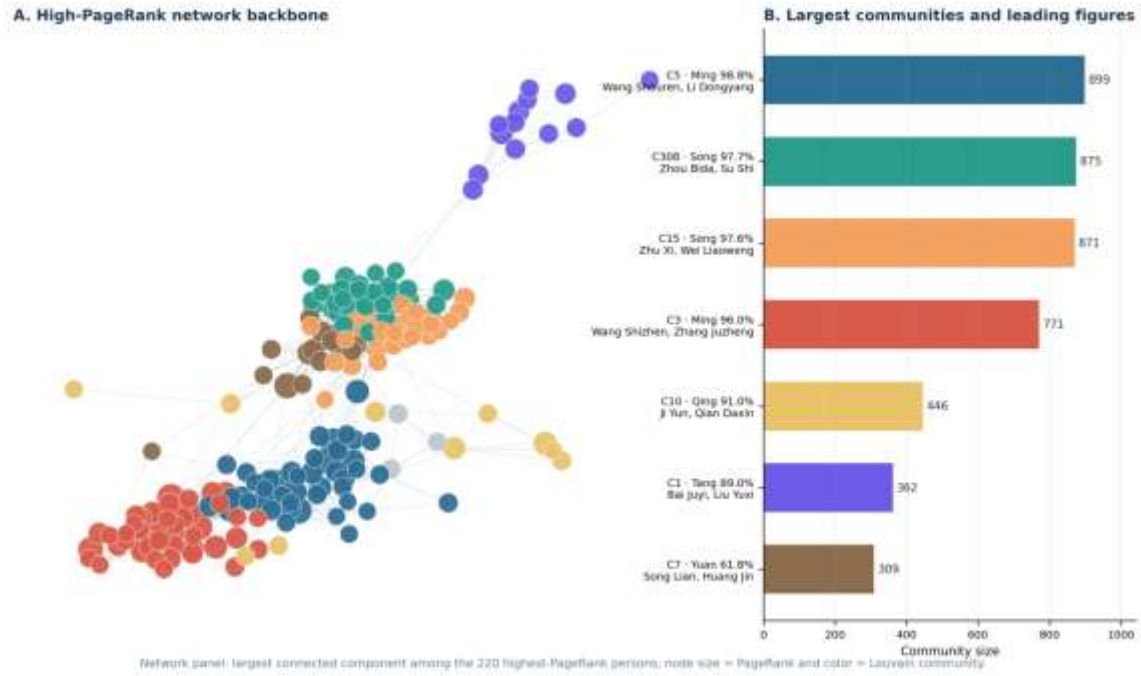

Figure 4. Community structure and dynastic composition in the high-information network backbone.
Source: Authors' analysis of the CBDB outputs.

4.4 CENTRALITY, PROMINENCE, AND CONVERGENT EVIDENCE

Weighted PageRank identifies Zhu Xi as the leading structural hub (PageRank = .01116, degree = 290, weighted strength = 1,997, approximate betweenness = .3215). He is followed by Zhou Bida (.00718), Su Shi (.00662), Ouyang Xiu (.00607), and Wang Anshi (.00451). The leading group is dominated by Song figures, with Ming figures such as Wang Shizhen, Zhang Juzheng, Wang Shouren, and Li Dongyang also appearing near the top. These rankings indicate prominence within the selected recorded multiplex network, not an ordinal judgment of cultural importance. The ten leading cases and their complementary centrality values are reported in Table 4.

KPS and PageRank are moderately associated (Spearman ρ = .4649, bootstrap 95% CI [.4421, .4867]). This

supports P1: heterogeneous archival visibility and network topology overlap, but neither is reducible to the other. KPS is highest for persons broadly represented across relation domains, whereas PageRank rewards recursive connectedness within the selected social-kinship graph. The distinction is analytically productive. Persons with high KPS but moderate PageRank may be richly documented across offices, texts, places, or events without occupying a central person-person network position; persons with high PageRank but lower KPS may be structurally central within a narrower relational record.

Table 3 reports the full rank-correlation matrix. The wider correlation matrix reinforces this interpretation. Degree and weighted strength are nearly redundant in rank order (ρ = .970), and both correlate strongly with PageRank (ρ = .887 and .893). PageRank and approximate betweenness correlate at .735, indicating substantial but incomplete overlap between recursive prominence and brokerage. KPS correlates more modestly with degree (.448), strength (.431), PageRank (.465), and betweenness (.391). Thus, KPS supplies a genuinely different dimension of evidence rather than a disguised centrality measure.

The alternative KPS weights produce strong but non-identical rankings within the exported top-1,000 table. Correlations with the equal-weight baseline are ρ = .938 for the coverage-focused score, ρ = .873 for the network-oriented score, and ρ = .923 for the documentation-oriented score. The result supports the use of KPS as a transparent prioritization device while showing that its exact ordering is not value-neutral. Figure 5D reports these bounded comparisons; full-database component exports would be required for a complete rank-replacement analysis.

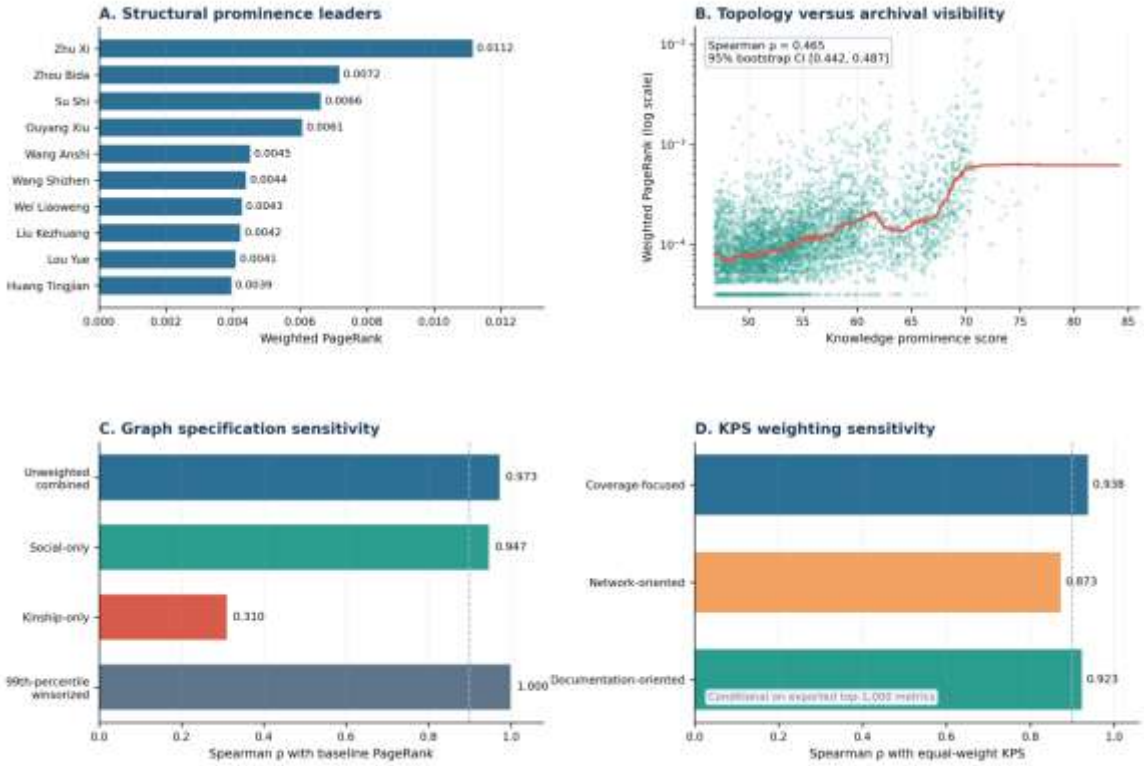

Figure 5. Structural prominence, archival visibility, graph robustness, and KPS weighting sensitivity. Source: Authors' analysis of the CBDB outputs.

Table 3. Spearman rank correlations among centrality and prominence measures

Measure	Degree	Strength	PageRank	Betweenness	KPS
Degree	1.000	0.970	0.887	0.629	0.448
Strength	0.970	1.000	0.893	0.619	0.431
PageRank	0.887	0.893	1.000	0.735	0.465
Betweenness	0.629	0.619	0.735	1.000	0.391
KPS	0.448	0.431	0.465	0.391	1.000

Note: N = 5,000. The PageRank–KPS bootstrap 95% confidence interval is [.442, .487].

Table 4. Leading persons by weighted PageRank

Person	Dynasty	Degree	Strength	PageRank	Betweenness	KPS
Zhu Xi	Song	290	1,997	0.01116	0.321	70.30
Zhou Bida	Song	251	1,214	0.00718	0.120	70.23
Su Shi	Song	200	1,171	0.00662	0.200	71.36
Ouyang Xiu	Song	179	1,034	0.00607	0.096	71.00
Wang Anshi	Song	192	679	0.00451	0.029	76.50
Wang Shizhen	Ming	171	678	0.00439	0.051	69.83
Wei Liaoweng	Song	154	706	0.00426	0.109	77.99
Liu Kezhuang	Song	151	682	0.00422	0.035	62.32
Lou Yue	Song	172	620	0.00408	0.022	62.15
Huang Tingjian	Song	118	720	0.00395	0.039	70.23

Note: Rankings are conditional on the KPS-selected 5,000-person graph. Betweenness is approximate (k = 200, seed = 42).

4.5 ROBUSTNESS AND LAYER DEPENDENCE

Table 5 reports the graph-specification comparisons. The baseline PageRank ordering is highly stable when edge weights are removed. The unweighted combined graph correlates with the baseline at ρ = .9729 and preserves 90% of the top 10, top 20, and top 50. Winsorizing weights at the 99th percentile produces almost identical ranks (ρ = .9997), with 90% top-10 overlap and complete overlap at the top 20 and top 50. These results support P3 and show that the principal ranking is not an artifact of a few extreme multiplicities.

The social-only graph also remains close to the baseline (ρ = .9467) and preserves every person in the baseline top 10, top 20, and top 50, although internal order can change. By contrast, the kinship-only graph correlates with the baseline at only ρ = .3101, with overlap of 10% in the top 10, 15% in the top 20, and 22% in the top 50. This strongly supports P2. The baseline combined graph is driven mainly by social-association structure, while kinship identifies a substantively different set of prominent persons.

The kinship result is not merely a weaker version of the social ranking. It changes the identity of the leading set and redirects attention from documented association and office-intellectual connectivity toward genealogical embeddedness. Accordingly, the combined graph should be described as association-dominant, while kinship should remain a separate interpretive layer rather than being absorbed into a single universal hierarchy (Fuller & Wang, 2021).

Table 5. Robustness and sensitivity analysis

Specification	Spearman rho	Top-10 overlap	Top-20 overlap	Top-50 overlap
Unweighted combined	0.973	0.900	0.900	0.900
Social-only	0.947	1.000	1.000	1.000
Kinship-only	0.310	0.100	0.150	0.220
99th-percentile winsorized	1.000	0.900	1.000	1.000

Note: Rank correlations and overlap rates are calculated against the weighted combined baseline PageRank ordering.

4.6 ANALYTICAL PROPOSITION ASSESSMENT

P1 is supported. PageRank and KPS are positively related (ρ = .465, 95% bootstrap CI [.442, .487]) but retain distinct leaders and only moderate rank convergence.

P2 is strongly supported. The social-only ranking remains close to the baseline (ρ = .947), whereas the kinship-only ranking diverges sharply (ρ = .310; top-50 overlap = 22%).

P3 is supported. Removing edge weights yields ρ = .973 with the baseline, and 99th-percentile winsorization yields ρ = 1.000; top-rank overlap remains high.

P4 is supported with qualification. Dynasty and place concentrations are clearly visible, but missing temporal and spatial fields, source survival, and database construction prevent population-level interpretation.

4.7 INTEGRATED DISCUSSION

4.7.1 WHAT THE FRAMEWORK REVEALS

But this historical factor is “definitely the most important” is not the point. It is the history of its recorded

prominence that can be broken down in various ways and compared. The recursive positioning in a social network of kinship (Weighted PageRank) emphasizes this dimension, as does the approximate betweenness, and the breadth of representation across the various relation domains is foregrounded by KPS. The moderate convergence suggests that they can contribute well to the cultural knowledge discovery via multiple views. If it is just once mention/record, it would be a combination of epistemic signal and would be difficult to understand how a person is being surfaced.

Song is historically a dominating category for the PageRank leaders but it is methodologically conditional. It may reflect close relationships among elite records in the archives, the extent to which the sources are coded in the CBDB, the extent of the coding and the nature of the intellectual and official links. The framework thus makes the ranking into a manageable list of cases for follow-up based on their sources, rather than a claim about dynastic cultural significance.

The social versus kinship result is particularly significant. A combined graph can pass the standard tests with high scores and have a single relation layer. The social-only ranking is a fairly decent approximation of the base ranking, so the combined PageRank should be used primarily as an indicator of association networks. The discovery agenda is different from the kinship layer, and hence the ranking of it. Multiplex sensitivity analysis shows this dependence before the construction of narratives by scholars.

4.7.2 THEORETICAL CONTRIBUTION

The study moves the competitive-intelligence mindset into digital humanities without taking into account market assumptions. It has a process model of scholarly intelligence as its theoretical contribution. Source intelligence sets the groundwork for determining whether evidence exists and if there is synergy within the release. Semantic integration maintains the identity of entities and the type of relations. Analytical intelligence: Analyzing and comparing coverage, topology, communities, and spatial patterns. Cultural discovery pinpoints candidate hubs, brokers, clusters and gaps. Ontology refinement and source acquisition are distillated by feedback. This is in line with the strategic focus on bringing information to action (Cavallo et al., 2021), yet reinterprets action as prioritizing, curating, and interpreting research.

The framework also introduces epistemic-governance layer for cultural knowledge graph. No limitations paragraph for provenances and missingness; provenances/non-provenances dictate which analyses are legitimate. High missingness of birth and death years, etc., is not reported because the model failed to grow over the years. It prevents this model from being claimed in the first place. But likewise, the metadata-checksum mismatch is stored as a consequence as well: because release identity is part of reproducibility. This orientation follows FAIR principles, which consider findability and accessibility as too little if it cannot be versioned, interpreted, and reused (Jacobsen et al., 2020).

4.7.3 METHODOLOGICAL CONTRIBUTION

Methodologically, a two-branch design is contributed to the paper explicitly. When using full-database profiling, all 659,593 persons are used, and the profiling gives the context necessary to understand the selected graph. A high information cohort that is bounded and transparent is then used for network analysis. Several large-data studies talk about the limits of computation but don't articulate what the implications of selection are. In this instance, each network claim is identified by the conditional operator "KPS: " and the entire database can be seen using coverage and composition tables.

The robustness design is more than just a minor set of parameter changes to an algorithm. It evaluates three different threats: weight dependence, outlier domination, and layer dependence. The first two threats are inconsequential for the PageRank ordering, the third is not. This pattern results in a more descriptive conclusion than a simple binary statement concluding that the model is sound. It reports to future researchers that the relation choice for the substantive relations is important but that computational implementation is stable.

The KPS is another methodological step forward because it is auditable, non-latent and can prevent an abundant relation domain from deciding a selection. The weighting diagnostic, however, reveals that there is no such thing as judgment-free weighting: the weighting scheme which is most network-based gives the greatest deviation from the baseline. In a substantive application, historians need therefore, to establish domain weights with respect to a given question and/or to compare different schemes.

The terms reliability and validity also need to correspond to the process used to create the data. The office, kinship, text, institution, event, and place counts are not reflective measures, and therefore are not suitable for Cronbach's alpha, composite reliability, AVE, and HTMT. Evidence is assessed using different operational definitions, convergence of ranks, partial correlation, alternative specifications, and differences that are understood in the context of history.

4.7.4 CONTEXTUAL AND SUBSTANTIVE IMPLICATIONS

The framework provides a scalable pathway from the large prosopography database to cases that focus on evidence, useful for historians of China. The relations that lead to recursive prominence can be studied in a PageRank leader, a possible bridge can be studied by an individual with a high betweenness, and an individual with a high KPS and a moderate PageRank can provide rich cross-domain documentation while being socially under-embedded. These are Discovery Prompts – not conclusions!

Library and cultural institution coverage measures can inform curation. Institution and event relations are very limited and place and temporal fields do not exist. If the domains are strategic in terms of importance, they can be used to identify the specific priorities in terms of authority control, source enrichment and spatial reconciliation. It is good to have semantic standards because they enable such gaps to be compared across collections (O'Neill & Stapleton, 2022).

Knowledge-graph developers need to do more than just count the nodes and edges in their knowledge graph. The following attributes of cultural graphs should be reported: entity class coverage, version provenance, connectedness, layer contribution, gender and source asymmetry, and stability of the ranking. The present framework extends rich domain modelling, which has been shown to be valuable for the Chinese context in the form of heritage graphs (Fan et al., 2023; Hou & Yuan, 2023; Liang et al., 2025), with an analytical audit.

4.7.5 LIMITATIONS AND FUTURE RESEARCH

CBDB is first and foremost a scholarly database that has been compiled from surviving sources and selected ones. It is not a census. All results are for recorded persons and relations. Source survival and coding practice affect dynasty and other distributions of place and gender.

Second, the network branch is chosen by KPS. The 5,000-person graph has been designed to be informative and therefore does not allow an estimate of the properties of the network for the whole person table. Comparisons with random, dynasty oriented, gender-aware and relation specific selection of KPS should be done in the future, or scaled graph systems used across the whole network.

Third, the undirected projection discards direction and role semantics. There are some social relationships and related kinship that are asymmetrical and type-specific. Centrality in the relation families should be reported on a typed, directed multiplex analysis, not by one projection alone.

Fourth, edge weights are multiplicities instead of frequencies of interaction or strength of ties. While the unweighted and winsorized results are consistent, source-aware weights should be able to differentiate independent attestations from duplicated, reciprocal or repeatedly coded records.

Fifth, there is an incomplete temporal information. Over 85% of persons lack birth years, and over 50% lack index year.Over 85% of the persons do not have birth years and over 50% do not have index years. Dynamic network analysis would demand a clearly delimited temporal subset, uncertainty time bounds for dates, and a clear choice analysis.

Last, the results provided by available analytical exports are not broken down by source family. In a direct source-bias study, the data from BIOG_TEXT_DATA would be joined with TEXT_CODES, the official histories, gazetteers, genealogies, collected works, and other source traditions would be classified, and then, the visibility and centrality would be re-computed with source-stratified specifications. No source-effect estimate is claimed here.

Seventh, the database is not representative, and the KPS-selected graph is even more unrepresentative, of women. This is a property of the surviving and coded record, plus the prioritization rule, and not an estimate of the participation of women in history. Future research should develop female-specific and relation-specific graphs and determine if this alters the visibility of the female.

5. CONCLUSION

In this study, a competitive intelligence-based knowledge graph framework of cultural knowledge discovery is proposed, and the performance of the proposed framework is verified using the China Biographical Database. The framework sees intelligence as an evidence-to-insight process that is accountable, not market-driven. It starts with release validation and coverage diagnosis, maintains heterogeneous cultural relations, splits the full-database description from the selected network analysis, and assesses graph conclusions for different specifications.

The empirical findings explain the importance of this governance. Although CBDB is extremely broad, it is not comprehensive and it has high temporal missingness. The PageRank and heterogeneous visibility to records in the high-information graph, with 5,000 users, are more similar but still demonstrate complementary notions of prominence. Rankings are robust against outlier treatment and weighting. Rankings are sensitive to whether social association or kinship is studied. The most plausible result is thus not a general ranking of the importance of those events for which historical sources exist. It's a clear representation of what's been seen and not seen, of what's in the structure and what isn't, of what depends on which layer and what doesn't, and of what evidence is missing and where. The gender diagnostic also reveals that, if the gender coded records were removed from the complete data set, only the evidence domain on the tasks would be reduced to 0.68% of the graph, while KPS values indicate that the prioritization remains sensitive to the weighting of the evidence domains.

The framework builds on digital humanities by introducing the concept of the analytical outcome including aspects of provenance, bias, robustness and interpretability. But its overarching message is straightforward: cultural knowledge graphs are scholarly intelligence systems only if they reflect not only what is linked, but how the linkage has been documented, what choices have been made in the modelling process, how discoveries are made, and where the evidence is lacking.

REFERENCES

1. Abu-Salih, B. (2021). Domain-specific knowledge graphs: A survey. Journal of Network and Computer Applications, 185, 103076. https://doi.org/10.1016/j.jnca.2021.103076

2. Ali, M., Berrendorf, M., Hoyt, C. T., Vermue, L., Galkin, M., Sharifzadeh, S., Fischer, A., Tresp, V., & Lehmann, J. (2022). Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12), 8825–8845. https://doi.org/10.1109/TPAMI.2021.3124805

3. Bao, Y. (2020). Competitive intelligence and its impact on innovations in tourism industry of China: An

empirical research. PLOS ONE, 15(7), e0236412. https://doi.org/10.1371/journal.pone.0236412
4. Bol, P. (2020). The visualization and analysis of historical space. Journal of Chinese History, 4(2), 511–519. https://doi.org/10.1017/jch.2020.22
5. Casillo, M., De Santo, M., Mosca, R., & Santaniello, D. (2023). Sharing the knowledge: Exploring cultural heritage through an ontology-based platform. Journal of Ambient Intelligence and Humanized Computing, 14, 12317–12327. https://doi.org/10.1007/s12652-023-04652-3
6. Cavallo, A., Sanasi, S., Ghezzi, A., & Rangone, A. (2021). Competitive intelligence and strategy formulation: Connecting the dots. Competitiveness Review, 31(2), 250–275. https://doi.org/10.1108/CR-01-2020-0009
7. Chen, C.-M., Chang, C., & Chen, Y.-T. (2023). A character social network relationship map tool to facilitate digital humanities research. Library Hi Tech, 41(2), 516–542. https://doi.org/10.1108/LHT-08-2020-0194
8. Chen, S., & Wang, H. (2022). China Biographical Database (CBDB): A relational database for prosopographical research of pre-modern China. Journal of Open Humanities Data, 8, Article 4, 1–6. https://doi.org/10.5334/johd.68
9. De Weerdt, H. (2020). Creating, linking, and analyzing Chinese and Korean datasets: Digital text annotation in MARKUS and COMPARATIVUS. Journal of Chinese History, 4(2), 519–527. https://doi.org/10.1017/jch.2020.23
10. Fan, T., Wang, H., & Hodel, T. (2023). CICHMKG: A large-scale and comprehensive Chinese intangible cultural heritage multimodal knowledge graph. Heritage Science, 11, Article 115. https://doi.org/10.1186/s40494-023-00927-2
11. Fuller, M., & Wang, H. (2021). Structuring, recording, and analyzing historical networks in the China Biographical Database. Journal of Historical Network Research, 5(1), 248–270. https://doi.org/10.25517/jhnr.v5i1.123
12. Guan, S., Cheng, X., Bai, L., Zhang, F., Li, Z., Zeng, Y., Jin, X., & Guo, J. (2023). What is event knowledge graph: A survey. IEEE Transactions on Knowledge and Data Engineering, 35(7), 7569–7589. https://doi.org/10.1109/TKDE.2022.3180362
13. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys, 54(4), Article 71, 1–37. https://doi.org/10.1145/3447772
14. Hou, Y., & Yuan, L. (2023). Building a knowledge graph of Chinese Kung Fu masters from heterogeneous bilingual data. Journal of Open Humanities Data, 9, Article 27, 1–12. https://doi.org/10.5334/johd.136
15. Hyvönen, E. (2020). Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web, 11(1), 187–193. https://doi.org/10.3233/SW-190386
16. Hyvönen, E. (2023). Digital humanities on the Semantic Web: Sampo model and portal series. Semantic Web, 14(4), 729–744. https://doi.org/10.3233/SW-223034
17. Ikkala, E., Hyvönen, E., Rantala, H., & Koho, M. (2022). Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web, 13(1), 69–84. https://doi.org/10.3233/SW-210428
18. Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., … Schultes, E. (2020). FAIR principles: Interpretations and implementation considerations. Data Intelligence, 2(1–2), 10–29. https://doi.org/10.1162/dint_r_00024
19. Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems, 33(2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843
20. Koho, M., Burrows, T., Hyvönen, E., Ikkala, E., Page, K., Ransom, L., Tuominen, J., Emery, D., Fraas, M., Heller, B., Lewis, D., Morrison, A., Porte, G., Thomson, E., Velios, A., & Wijsman, H. (2022). Harmonizing and publishing heterogeneous pre-modern manuscript metadata as linked open data. Journal of the Association for Information Science and Technology, 73(2), 240–257. https://doi.org/10.1002/asi.24499
21. Krabina, B. (2023). Building a knowledge graph for the history of Vienna with Semantic MediaWiki. Journal of Web Semantics, 76, 100771. https://doi.org/10.1016/j.websem.2022.100771
22. Liang, Y., Xie, B., Tan, W., & Zhang, Q. (2025). Ontology-based construction of embroidery intangible cultural heritage knowledge graph: A case study of Qingyang sachets. PLOS ONE, 20(1), e0317447. https://doi.org/10.1371/journal.pone.0317447
23. Maluleka, M. L., & Chummun, B. Z. (2023). Competitive intelligence and strategy implementation: Critical examination of present literature review. South African Journal of Information Management, 25(1), Article a1610, 1–12. https://doi.org/10.4102/sajim.v25i1.1610
24. Maungwa, T., & Laughton, P. (2023). The use of theories in competitive intelligence: A systematic literature review. Journal of Intelligence Studies in Business, 13(2), 43–60. https://doi.org/10.37380/jisib.v13i2.1083
25. O'Neill, B., & Stapleton, L. (2022). Digital cultural heritage standards: From silo to semantic web. AI & Society, 37(3), 891–903. https://doi.org/10.1007/s00146-021-01371-1
26. Ranjan, J., & Foropon, C. (2021). Big data analytics in building the competitive intelligence of organizations. International Journal of Information Management, 56, 102231. https://doi.org/10.1016/j.ijinfomgt.2020.102231
27. Renzi, G., Rinaldi, A. M., Russo, C., & Tommasino, C. (2023). A storytelling framework based on multimedia knowledge graph using linked open data and deep neural networks. Multimedia Tools and

Applications, 82, 31625–31639. https://doi.org/10.1007/s11042-023-14398-x
28. Rossi, A., Barbosa, D., Firmani, D., Matinata, A., & Merialdo, P. (2021). Knowledge graph embedding for link prediction: A comparative analysis. ACM Transactions on Knowledge Discovery from Data, 15(2), 1–49. https://doi.org/10.1145/3424672
29. Tsui, L. H., & Wang, H. (2020). Harvesting big biographical data for Chinese history: The China Biographical Database (CBDB). Journal of Chinese History, 4(2), 505–511. https://doi.org/10.1017/jch.2020.21
30. Wen, F., Wang, E. H., & Hout, M. (2024). Social mobility in the Tang Dynasty as the Imperial Examination rose and aristocratic family pedigree declined, 618–907 CE. Proceedings of the National Academy of Sciences, 121(4), e2305564121. https://doi.org/10.1073/pnas.2305564121