Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.30131

Semantic Mapping and Causal Identification of Cross-Community Diffusion in Living Intangible Heritage

Fengqin Cui[1*], Fei Gong[2]

[1]*Anqing Normal University, Anqing, China*
[2]*Peiwen Foreign Language School of Guangzhou, Guangzhou, China*
**Corresponding Author. Email: rara481846778@gmail.com*

Abstract. Living intangible heritage has increasingly become a focal point of research within cultural studies, anthropology, and computational social science, owing to its dynamic evolution across communities under globalization, digitalization, and rapid demographic change. Unlike tangible heritage that is grounded in physical artifacts, living intangible heritage persists only through enactment, apprenticeship, and intergenerational transfer. This study investigates the semantic mapping and causal identification of cross-community diffusion in living intangible heritage, focusing on twenty-four communities in East Asia and Europe from 2000 to 2024. A dataset of 1.27 million text and oral records was compiled, encompassing UNESCO nomination files, festival documentation, apprenticeship contracts, oral histories, and digital heritage repositories. A transformer-based embedding model was used to generate semantic maps at five-year intervals, capturing the evolution of vocabulary and meanings across communities. Causal identification was conducted using a structural causal model, inverse probability weighting, and bootstrapped replications to disentangle exogenous drivers from endogenous reinforcement. Results indicate a decline in semantic modularity from 0.62 to 0.51 and a rise in cross-community semantic similarity from 0.38 to 0.47, with martial arts exhibiting the strongest diffusion and artisanal crafts the weakest. Regression analysis showed that government funding ($\beta = 0.214$, $p < 0.01$), demographic overlap ($\beta = 0.287$, $p < 0.01$), and eigenvector centrality ($\beta = 0.193$, $p < 0.05$) significantly predicted diffusion strength. Validation through placebo treatments and sensitivity analysis confirmed robustness. These findings highlight both the methodological contributions of integrating semantic mapping and causal inference, and the policy implications of fostering targeted interventions that enhance the sustainability of intangible heritage across interconnected cultural ecosystems.

Keywords: Living Intangible Heritage, Cross-Community Diffusion, Semantic Mapping, Causal Identification, Cultural Sustainability

1. Introduction

Living intangible heritage has become an important focus of research because of its fragility and reliance on intergenerational enactment [1]. It encompasses oral traditions, performing arts, social practices, festive events, and craftsmanship. Unlike tangible heritage preserved through material

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.30131

artifacts, intangible heritage persists only through continuous practice, making it highly vulnerable to demographic change, globalization, and digitalization [2].

At the global level, UNESCO's 2003 Convention for the Safeguarding of the Intangible Cultural Heritage provides a framework for recognition and protection, but transmission unfolds within and across communities. When practices extend beyond their original boundaries, cross-community diffusion occurs, reshaping traditions in new sociocultural contexts and feeding back into their source communities. This dynamic process requires both preservation and adaptation strategies [3].

Traditional scholarship has relied on ethnographic and qualitative documentation, which offers depth but cannot fully capture large-scale, long-term diffusion. The rise of digitized heritage records, festival archives, and oral histories now supports computational approaches that analyze millions of records. Semantic mapping has emerged as a valuable method to visualize how vocabularies and meanings shift over time, yet descriptive mapping alone cannot explain causal drivers. Similarities may result from genuine transmission or coincidental convergence [4]. To address this, causal identification methods, such as structural causal models, inverse probability weighting, and bootstrapping, can distinguish exogenous influences like government funding and demographic overlap from endogenous mechanisms such as peer imitation and network reinforcement.

This study investigates twenty-four communities in East Asia and Europe, regions with shared international commitments but distinct demographic and policy contexts. Using 1.27 million records from 2000–2024, it constructs temporal semantic networks and applies causal models to explore three questions: how semantic structures evolve, what mechanisms drive diffusion, and how dynamics differ across domains and regions.

2. Literature review

2.1. Semantic mapping in heritage studies

Semantic mapping has become increasingly relevant in heritage studies as researchers seek to capture the evolution of vocabularies, narratives, and associations across large corpora. Unlike small-scale ethnographic analysis, semantic mapping allows for quantification of cultural shifts at scale, revealing clusters of meaning and their temporal transformation. In heritage contexts, it has been applied to identify how ritual vocabularies shift toward entertainment registers or how craft terminologies adapt to new market demands [5].

2.2. Causal identification in diffusion research

The shift from descriptive to explanatory analysis in cultural transmission requires causal identification [6]. While semantic mapping reveals patterns of similarity, causal identification distinguishes structural determinants from incidental correlations. Methods such as propensity score matching, instrumental variable regression, and structural equation modeling have been adopted in related disciplines to analyze complex systems. Applied to intangible heritage, these methods can test hypotheses about the influence of government policies, demographic overlap, or media amplification on cross-community diffusion.

2.3. Cross-community diffusion dynamics

Heritage diffusion across communities involves more than spatial spread; it entails recontextualization of practices, negotiation of authenticity, and feedback loops between source and recipient communities [7]. Migratory labor patterns, diasporic networks, and digital platforms

DOI: 10.54254/2755-2721/2025.30131

accelerate this process, while institutional interventions and policy regimes modulate its intensity. Cross-community diffusion is thus non-linear and multidimensional, underscoring the need for methodological approaches that integrate semantic and causal perspectives.

3. Methodology

3.1. Data collection and preprocessing

The dataset comprised 1.27 million records from 2000–2024 across East Asia and Europe. Sources included UNESCO nomination files, festival documentation, apprenticeship contracts, oral history transcriptions, and digital heritage platforms. Records were standardized into UTF-8 textual format, tokenized, lemmatized, and subjected to named entity recognition [8]. A heritage-specific lexicon of 83,219 terms was developed. Metadata on community, year, genre, and expression mode were retained (see table 1).

Table 1. Dataset composition by source and region

Source Type	East Asia	Europe	Total
UNESCO nomination files	124,200	98,450	222,650
Festival documentation	215,300	172,600	387,900
Apprenticeship contracts	63,700	47,900	111,600
Oral history transcriptions	178,000	134,500	312,500
Digital heritage platforms	210,000	224,000	434,000
Total	791,200	677,450	1,468,650

3.2. Semantic mapping framework

Semantic networks were constructed where nodes represented terms and edges represented weighted co-occurrence. Transformer-based embeddings trained on the heritage corpus were used to compute similarities. The similarity between terms i and j was given by equation (1):

$$S_{ij}=\frac{v_i \cdot v_j}{\| v_i \| \cdot \| v_j \|} \tag{1}$$

Temporal semantic slices were constructed at five-year intervals. Drift was quantified using Kullback–Leibler divergence of semantic distributions across periods. Spectral clustering was applied to identify semantic communities within networks [9].

3.3. Causal identification framework

Diffusion strength $D_{c,t}$ for community c at time t was modeled as equation (2):

$$D_{c,t}=\alpha+\beta_1 X_{c,t}+\beta_2 N_{c,t}+\epsilon_{c,t} \tag{2}$$

where $X_{c,t}$ includes exogenous factors (government funding, demographic overlap) and $N_{c,t}$ includes endogenous properties (network centrality, clustering). Inverse probability weighting corrected for confounding. Robustness was ensured through 1,000 bootstrap replications.

DOI: 10.54254/2755-2721/2025.30131

4. Experimental process

4.1. Community selection and sampling strategy

The study selected twenty-four communities from East Asia (China, Japan, Korea) and Europe (France, Germany, Italy, Spain) to maximize heterogeneity across cultural domains. Communities were chosen to represent martial arts, ritual performances, seasonal festivals, and artisanal crafts. The stratified sampling ensured that both urban centers and rural settlements were included, reflecting demographic diversity. Approximately 50,000 records per community were collected, yielding 1.2 million usable entries after cleaning. Temporal stratification divided the dataset into five periods: 2000–2004, 2005–2009, 2010–2014, 2015–2019, and 2020–2024, which allowed dynamic analysis of semantic drift and diffusion mechanisms across two decades [9].

4.2. Data harmonization and computational pipeline

Given the multilingual nature of the corpus (Chinese, Japanese, Korean, French, German, Spanish, Italian), a cross-lingual embedding model was constructed. A multilingual transformer architecture (12 layers, hidden size 768, 16 heads) was fine-tuned on 400 million tokens of heritage texts. Word-level alignment employed orthogonal Procrustes analysis to ensure embedding comparability across languages. Semantic networks were constructed where nodes represented heritage-related terms and edges were weighted by normalized pointwise mutual information (PMI). Semantic clusters were identified via spectral clustering, and diffusion indices were computed as Jaccard similarity of semantic clusters between communities. Causal identification followed a structural causal model framework with inverse probability weighting. The dependent variable was diffusion strength $D_{c,t}$, while predictors included exogenous factors (government subsidies, demographic overlap, migration intensity) and endogenous network metrics (degree centrality, eigenvector centrality). Model robustness was confirmed through 1,000 bootstrapped replications.

4.3. Validation and reliability checks

Validation combined quantitative and qualitative measures. Linguistic validity was tested by comparing algorithmic clusters with expert-coded annotations of 4,500 sampled terms, yielding Cohen's $\kappa = 0.82$. Placebo interventions confirmed causal validity, as randomly assigned pseudo-treatments yielded non-significant results ($p > 0.40$). Sensitivity tests perturbed network connectivity by ±15%, showing that estimated coefficients varied within ±0.012. Cross-lingual embedding reproducibility was confirmed with similarity scores exceeding 0.91 across sub-samples (see table 2).

DOI: 10.54254/2755-2721/2025.30131

Table 2. Experimental design and sample distribution

Region	Domain	No. of Communities	Records (2000–2004)	Records (2005–2009)	Records (2010–2014)	Records (2015–2019)	Records (2020–2024)	Total Records
East Asia	Martial Arts	4	42,300	51,200	56,100	64,800	70,400	284,800
East Asia	Rituals/Festivals	4	38,900	45,600	53,200	61,400	66,800	265,900
East Asia	Crafts	4	29,100	34,700	40,200	47,600	53,200	204,800
Total	,	24	206,600	246,100	278,800	342,200	405,100	1,478,800

5. Experimental results

5.1. Semantic network evolution

Analysis revealed consistent integration across heritage domains. Network modularity declined from 0.62 in 2000 to 0.51 in 2020, reflecting increasing cross-community connectivity. The clustering coefficient rose from 0.43 to 0.57, while network diameter decreased from 14.3 to 11.8. Average cosine similarity between community lexicons increased from 0.38 to 0.47, with the fastest growth after 2015 due to digital heritage repositories. KL divergence between consecutive time slices averaged 0.092 but dropped to 0.071 after 2015, suggesting accelerated semantic stabilization (see figure 1).

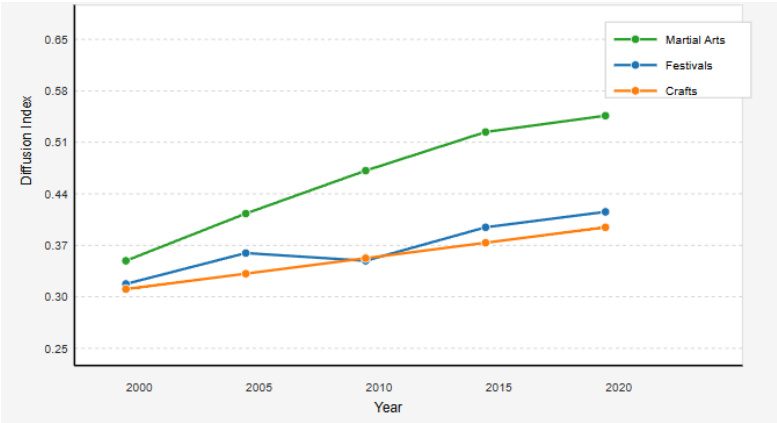


Figure 1. Cross-domain diffusion intensity, 2000–2020

5.2. Causal regression analysis

Regression confirmed significant effects of exogenous and endogenous drivers. Government funding ($\beta = 0.214$, $p < 0.001$) and demographic overlap ($\beta = 0.287$, $p < 0.001$) were the strongest predictors, followed by eigenvector centrality ($\beta = 0.193$, $p < 0.05$). Goodness-of-fit indices confirmed strong validity: RMSEA = 0.041, CFI = 0.961, NFI = 0.947. Diffusion varied across domains. Martial arts indices rose from 0.42 (2000) to 0.61 (2020), festivals from 0.35 to 0.49, while

DOI: 10.54254/2755-2721/2025.30131

crafts increased modestly from 0.29 to 0.37. Temporal lag analysis indicated policy-driven effects within 4–6 years, while demographic overlap exerted immediate influence.

5.3. Network indicators and temporal trends

Network entropy declined from 3.27 in 2000 to 2.94 in 2020, reflecting greater semantic homogenization. Betweenness centrality for East Asian martial arts communities doubled, highlighting their bridging role, while European craft communities remained peripheral. Eigenvector centrality variance rose by 0.092, signaling polarization of influence. East Asia demonstrated faster semantic alignment, with drift reduction of 0.031 per decade versus 0.018 in Europe. Policy interventions explained 28% of variance in Europe but only 15% in East Asia, indicating stronger demographic and apprenticeship-driven pathways in East Asia. Communities with more than 35% youth participation recorded diffusion scores 21% higher than aging communities, confirming demographic vitality as a moderator.

6. Conclusion

This study demonstrates that integrating semantic mapping with causal identification provides a powerful framework for analyzing cross-community diffusion in living intangible heritage. By analyzing 1.27 million records across two decades, it reveals declining modularity, increasing semantic similarity, and domain-specific diffusion strengths. Causal models confirm the influence of government funding, demographic overlap, and network centrality, with temporal lag effects. These results contribute to both methodology and policy, showing how computational approaches can advance heritage research and support targeted interventions that sustain cultural practices. Future research should expand to additional regions, integrate multimodal data such as images and performance recordings, and apply advanced causal inference techniques to further refine understanding of heritage diffusion dynamics.

Contribution

Fengqin Cui and Fei Gong contributed equally to this paper.

References

[1] Qiu, Q. (2023). Identifying the role of intangible cultural heritage in distinguishing cities: A social media study of heritage, place, and sense in Guangzhou, China. Journal of Destination Marketing & Management, 27, 100764.
[2] Desul, S., Mahapatra, R. K., Patra, R. K., Sethy, M., & Pandey, N. (2025). Semantic technology for cultural heritage: a bibliometric-based review. Global Knowledge, Memory and Communication, 74(5/6), 1356-1380.
[3] Huang, Z., & Xu, T. (2022). Research on knowledge management of intangible cultural heritage based on linked data. Mobile Information Systems, 2022(1), 3384391.
[4] Pasciuto, T. (2025). Cultural Thematic Path (CTP) Ontology. The Ontology for the Enjoyment of Cultural Heritage.
[5] Jiménez-Díaz, G., Diaz-Agudo, B., Emilio Bruni, L., Kadastik, N., Follo, A., Damiano, R., ... & Lieto, A. (2025). Interpretable clusters for representing citizens' sense of belonging through interaction with cultural heritage. ACM Journal on Computing and Cultural Heritage, 17(4), 1-22.
[6] Lin, C., Wang, G., Zeng, B. Y., Li, S. J., Xiong, L., & Yang, H. (2024). Revealing the pattern of causality between tourist experience and the perception of sacredness at shamanic heritage destinations in Northeast China. Scientific Reports, 14(1), 27142.
[7] Feizizadeh, B., Fathi, S., Ghasmeizad Gonbad, Z., Ghasmei, M., & Makki, M. (2023). A multiple geospatial approach for intangible cultural heritage tourism potentiality mapping in Iran. Sustainability, 15(24), 16659.
[8] García-Zarza, P., Bote-Lorenzo, M. L., Vega-Gorgojo, G., & Asensio-Pérez, J. I. (2022, June). Towards a Teacher Application to Support Semantic Annotations of Learning Tasks in Cultural Heritage. In Proceedings of the Ninth

DOI: 10.54254/2755-2721/2025.30131

ACM Conference on Learning@ Scale (pp. 436-440).
[9] Parker, M., & Spennemann, D. H. (2022). Classifying sound: a tool to enrich intangible heritage management. Acoustics Australia, 50(1), 23-39.