Paths of A Million People: Extracting Life Trajectories from Wikipedia

Ying Zhang*, Xiaofeng Li*, Zhaoyang Liu, and Haipeng Zhang†

章颖*、李笑风*、刘昭阳、张海鹏†

ShanghaiTech University

上海科技大学

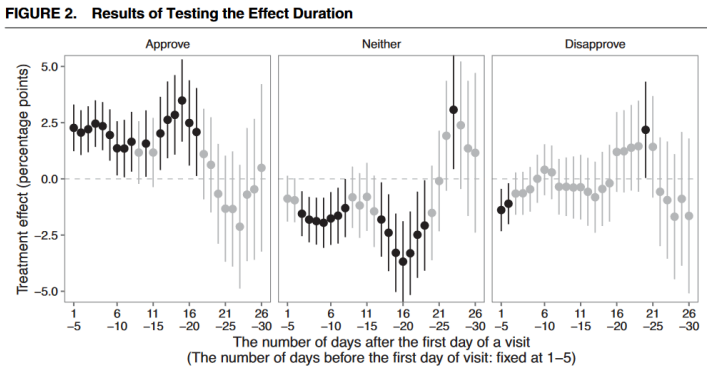
June 24, 2025

Paper Code Dataset

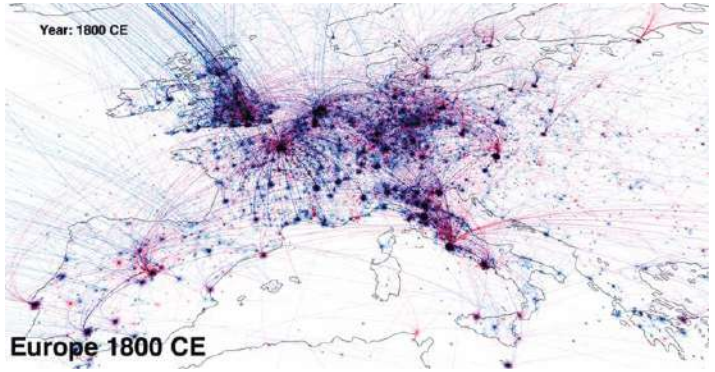
Our Recent Work Me

Motivation

- Life trajectories of notable individuals have crucial implications



Politics
(Goldsmith et al., 2021)



Cultural History
(Schich et al., 2014)

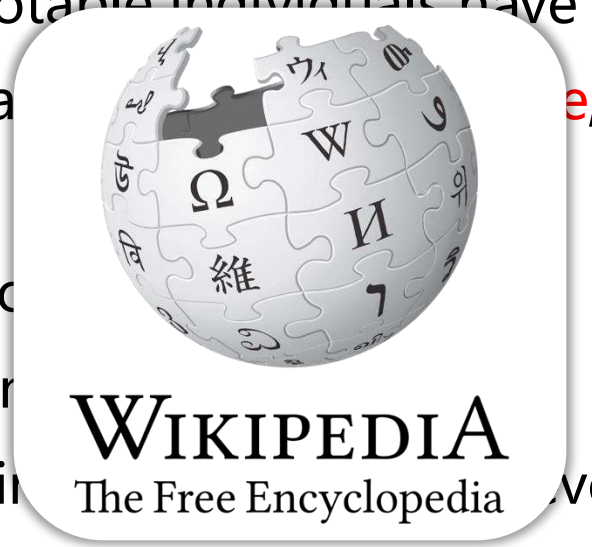
Motivation

- Life trajectories of notable individuals have crucial implications
- Trajectory data is scarce in terms of volume, density, and inter-person interactions
  - Usually <10 k footprints on record
  - Only a few-dozen trajectory types

Intermediate points unlock network-level analysis

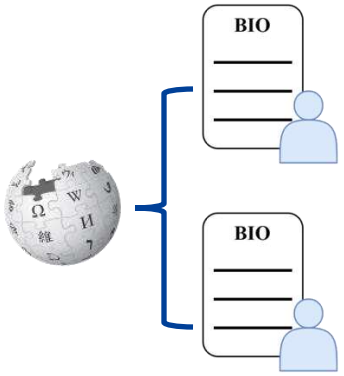
Motivation

- Life trajectories of notable individuals have crucial implications
- Trajectory data is sca e, density, and inter-person interactions
  - Usually <10 k foc
  - Only a few-dozer
  - 💡 Intermediate poi vel analysis
- Wikipedia has millions of biography pages, with abundant trajectory information



Problem Statement

- Life trajectory is consisted of three elements *(Person, Time, Location)*
- Formulation: Given the context, extract life trajectories by classifying the relevant triplets



Bob Hayes represented the USA in the **1964** Summer Olympics in **Tokyo**.

(**Bob**, **1964**, **Tokyo**) ✓

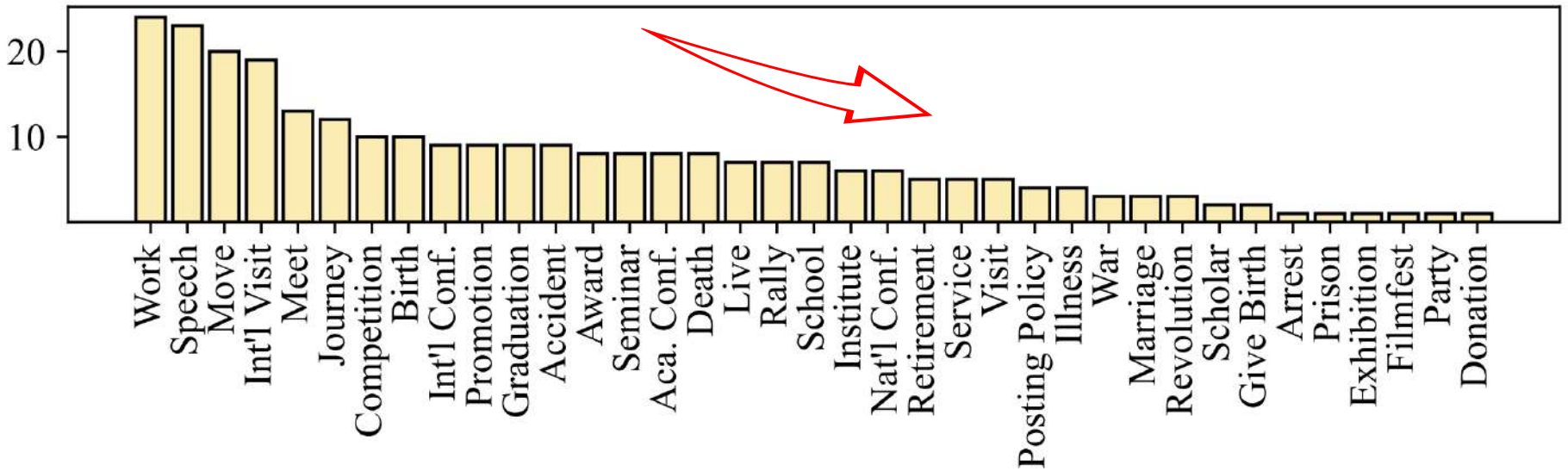
Janusz Symonides was born in Brest in **1938** and graduated from high school in **Toruń**.

(**Janusz**, **1938**, **Toruń**) ✗

Triplet: (PERSON TIME LOCATION)

Challenges

- More than **35** types are observed in just **10** random biographies
- Total **1,930,519** biography pages on Wikipedia
- How to generalize to **long-tail** data?



Similarity vs. Dissimilarity

Similar Pattern

① **Bob Hayes** represented the USA in the **1964** Summer Olympics in **Tokyo**. ✓

PERSON TIME LOCATION

② **Mark Nichols** stood for Canada in the **2022** Winter Olympics in **Beijing**. ✓

Different Pattern

③ **Janusz Symonides** was born in Brest in **1938** and graduated from high school in **Toruń**. ✗

④ **Bob Hayes** represented the **USA** in the **1964** Summer Olympics in Tokyo. ✗

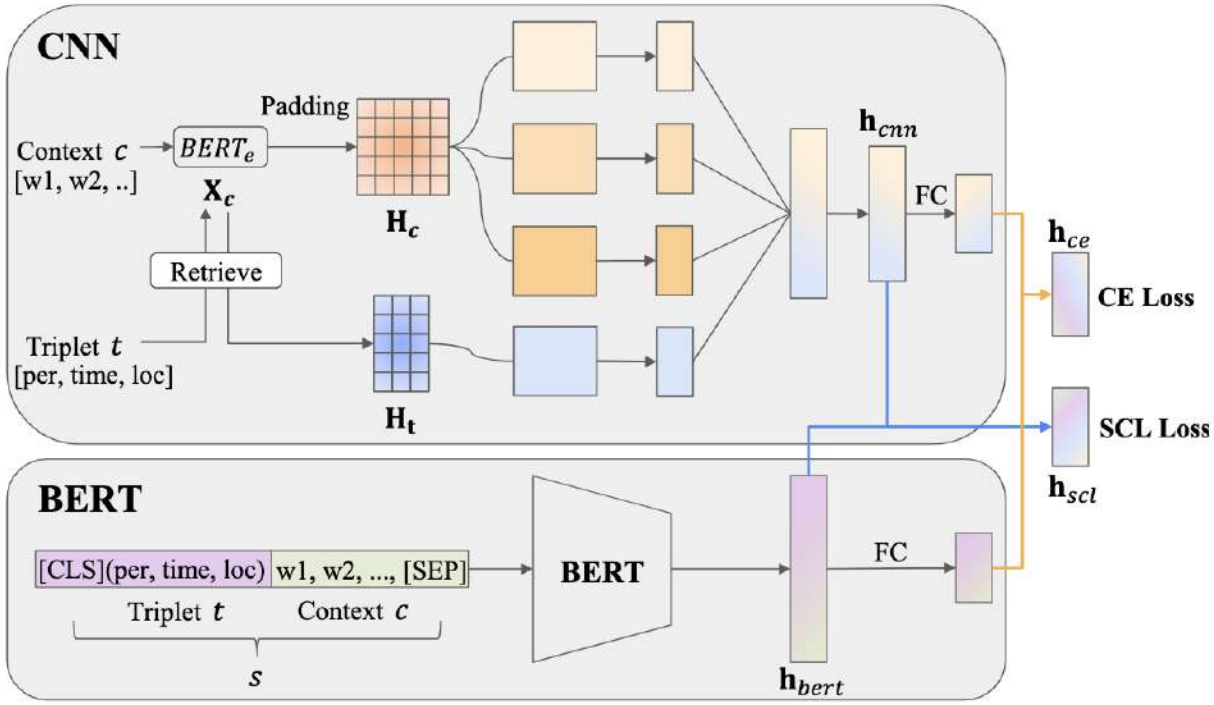
More samples in the wild …

COSMOS

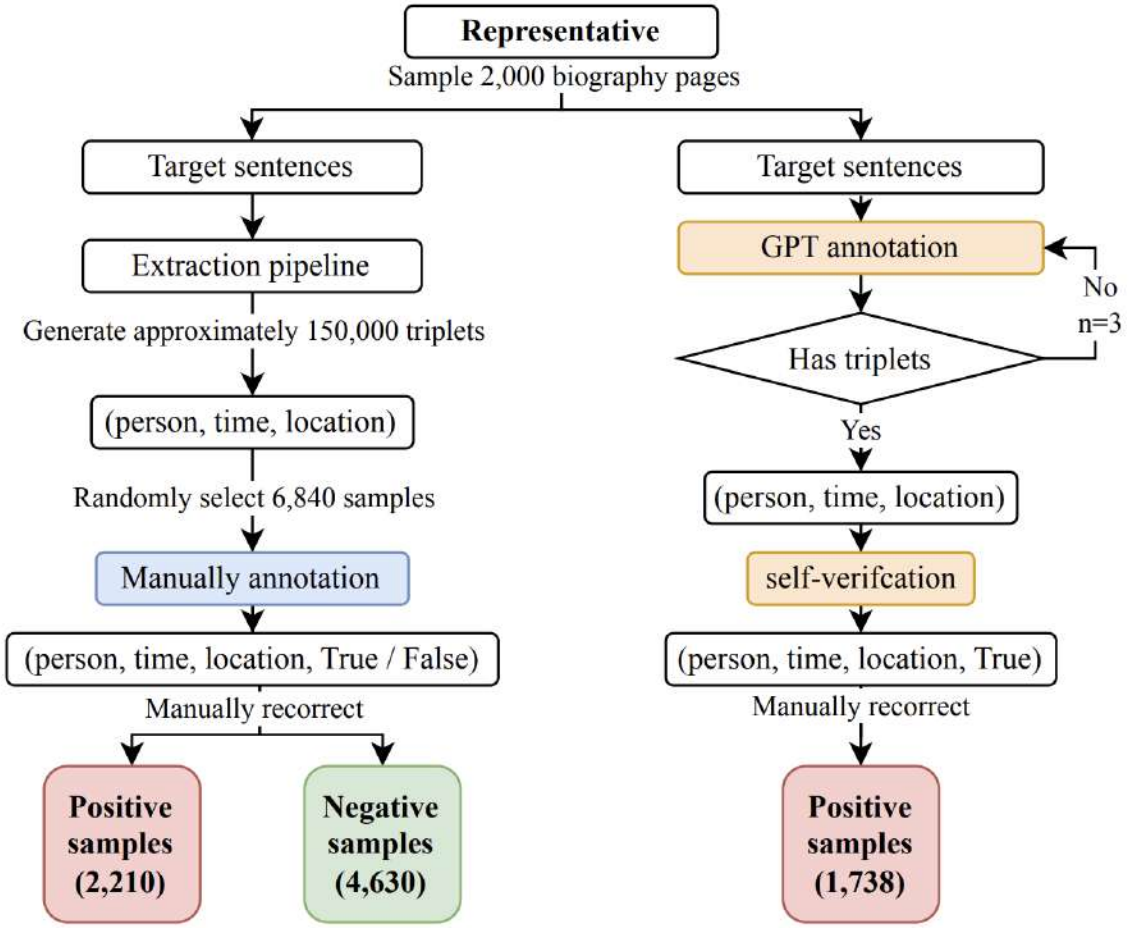
- Classification: Given *(Person, Time, Location)* t and its context p

$$f : \{t, p, \Theta\} \rightarrow y$$

- Use contrastive-learning to capture intra-sample relation
- Use semi-supervised learning to expand the training data during model training



Dataset



1. Develop a preprocess tool to mine candidate triplets
2. Construct a benchmark dataset
   - Representative -> Accuracy
   - Regular -> Coverage

Experimental Result

- Compared to seven baselines

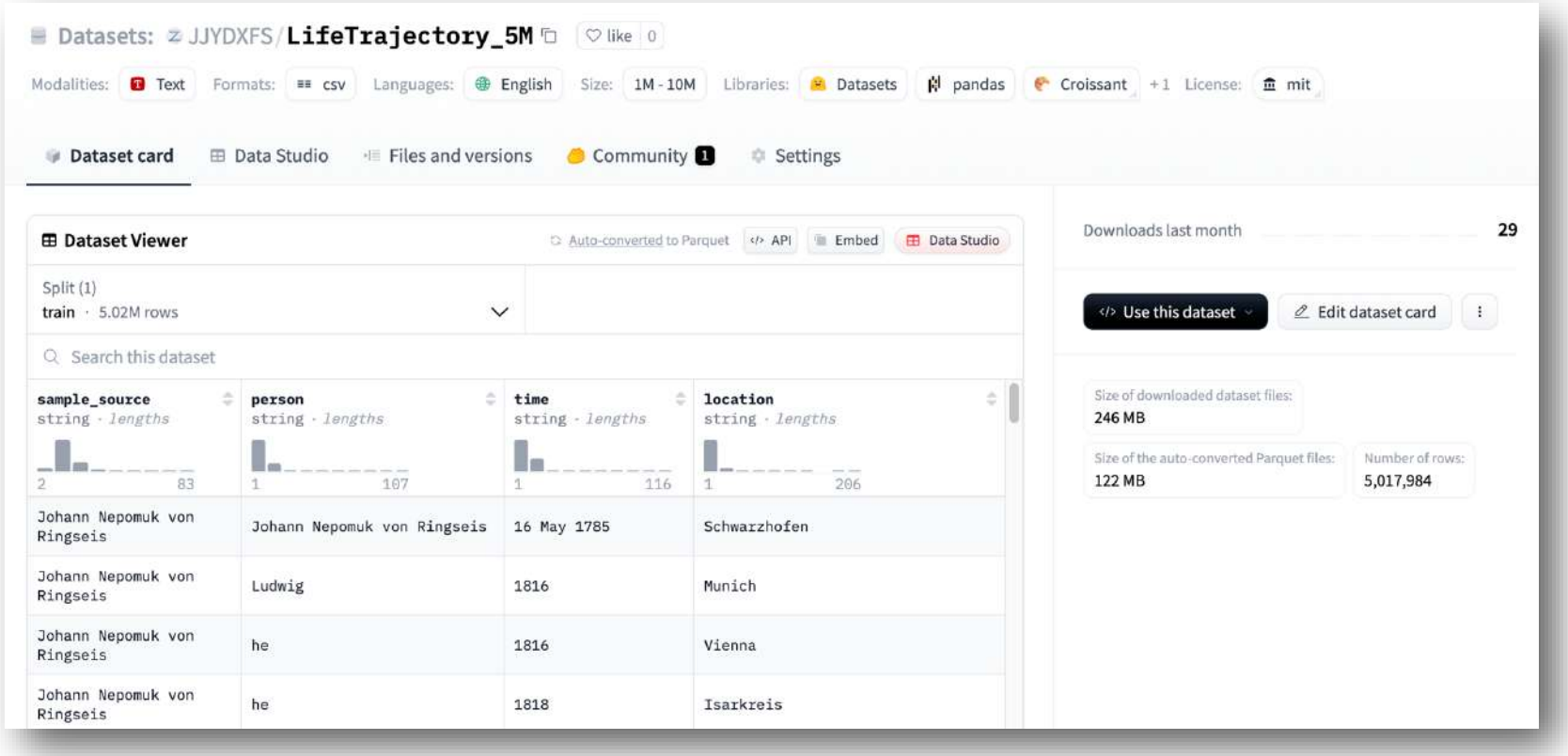
COSMOS achieves the best overall performance

COSMOS exhibits better generalization ability

	Representative				Representative$_m$				Representative$_g$	Regular	
	Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)	R (%)	R (%)	Avg-R (std)
GPT-3.5	63.99	56.53	95.12*	70.91	55.00	41.48	91.39*	57.06	100.00*	92.33*	0.9126 ± 0.0716
LR (TFIDF)	74.47	75.45	66.24	70.55	75.67	62.62	63.64	63.13	69.64	44.52	0.4262 ± 0.1751
CeleTrip	82.55	81.77	80.05	80.90	81.31	70.26	74.33	72.24	87.54	60.94	0.5614 ± 0.2351
Bi-LSTM	84.38	81.38	85.77	83.52	81.94	69.66	79.37	74.20	94.16	75.18	0.7549 ± 0.2031
CNN	84.42	**84.91**	80.55	82.67	82.62	**74.08**	72.10	73.08	91.63	63.50	0.6344 ± 0.2111
BERT	84.65	80.10	**88.80**	84.23	82.08	68.39	**84.12**	75.44	94.94	81.02	0.8304 ± 0.1398
RoBERTa	86.09	82.88	88.04	85.38	83.68	71.94	82.19	76.73	**95.71**	77.00	0.7389 ± 0.1583
COSMOS	**86.79**	84.41	87.54	**85.95**	**84.61**	**74.08**	81.45	**77.59**	95.52	**82.11**	0.8169 ± 0.0906

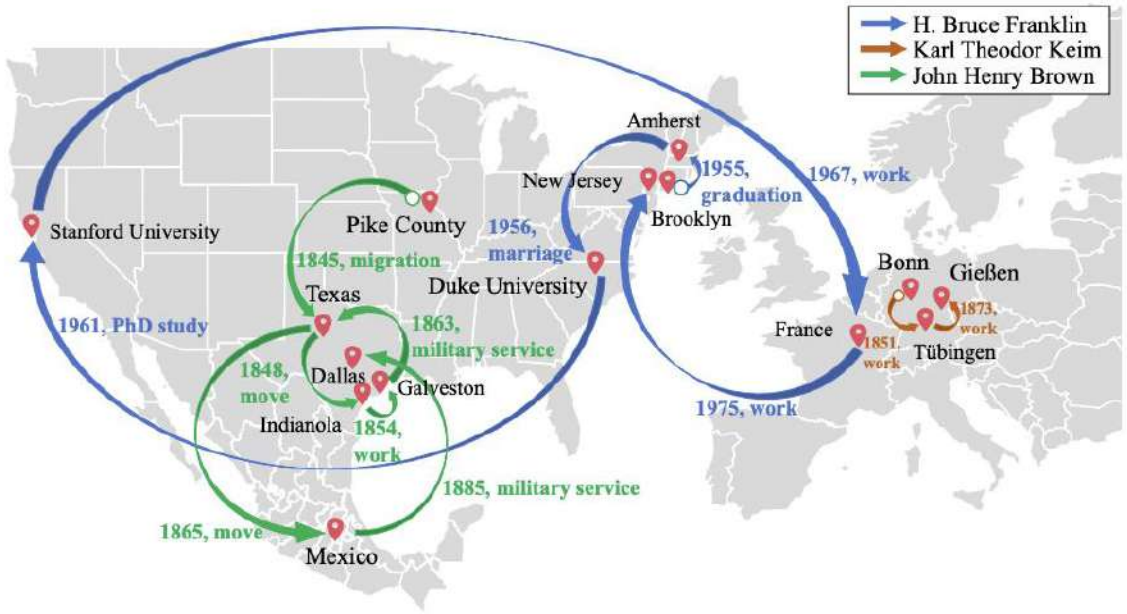
Millions of Extracted Life Trajectories

- From experiments to practice
- From 1,930,519 biographies to 5,017,984 life trajectory triplets

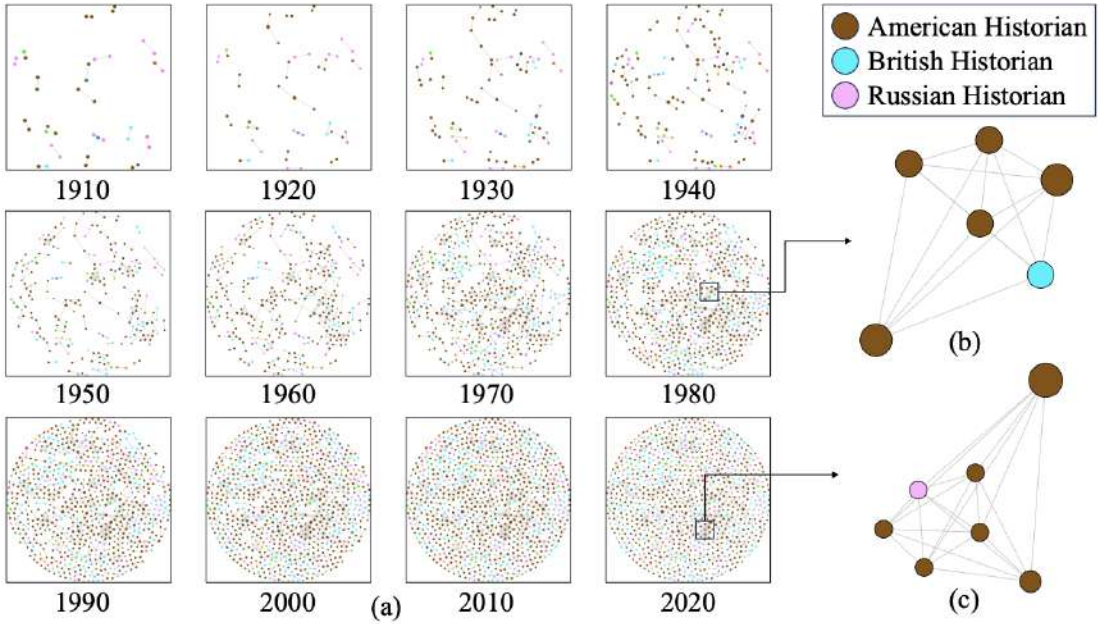


Life Trajectory of Historians

- 20,786 trajectory triplets of 8,272 historians
- Types: born / die / move / study / teach / marry / graduate / work …



Individual-level Trajectory Visualization

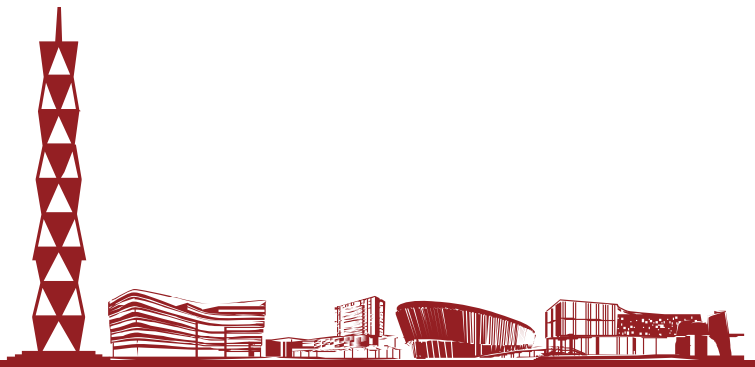


Group-level Trajectory Interaction

Summary

- Propose a **task** of life trajectory extraction, and our framework, **COSMOS**, for this task
- Release our **labeled dataset**, **million-level extracted trajectories** and open-source our **framework**

Paper Code Dataset



立志成才报国裕民