scan QR code
for more details

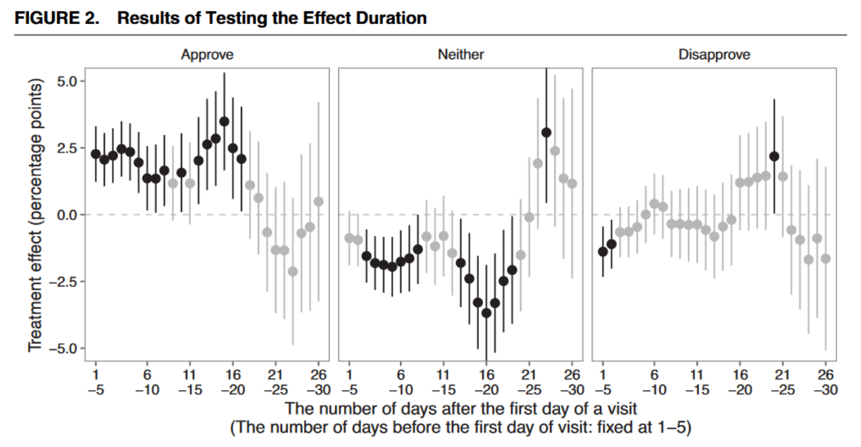
Paths of A Million People: Extracting Life Trajectories from Wikipedia

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

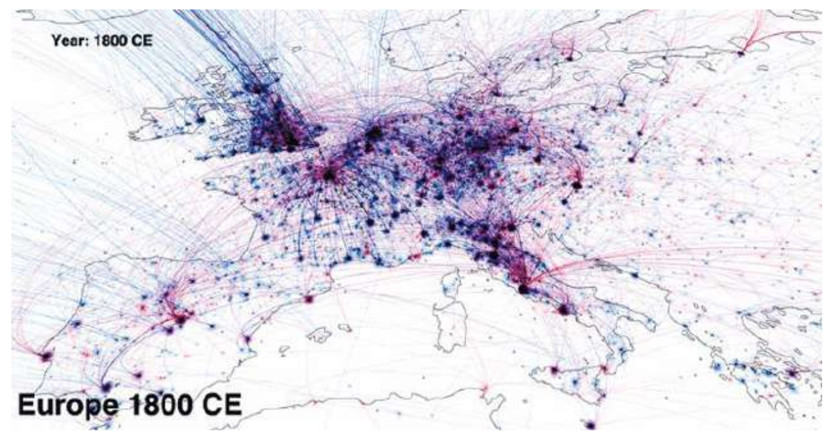
ShanghaiTech University

Motivation

Life trajectories of notable people have been studied to pinpoint the times and places of significant events such as birth, death, education, and battles. However, the scarcity of trajectory data in terms of **volume, density, and inter-person interactions**, limits relevant studies from being comprehensive and interactive.



Analysis of politicians' trajectory
(Goldsmith et al., 2021)



Birth and Death Places of Cultural Figures
(Schich et al., 2014)

We need a comprehensive trajectory dataset!

Related Work

Existing Rule-based Extraction

× Use **predefined** semantic roles from FrameNet
× Only considers 29 frames "related to movements"
× **Low Recall**

Supervised Learning Method

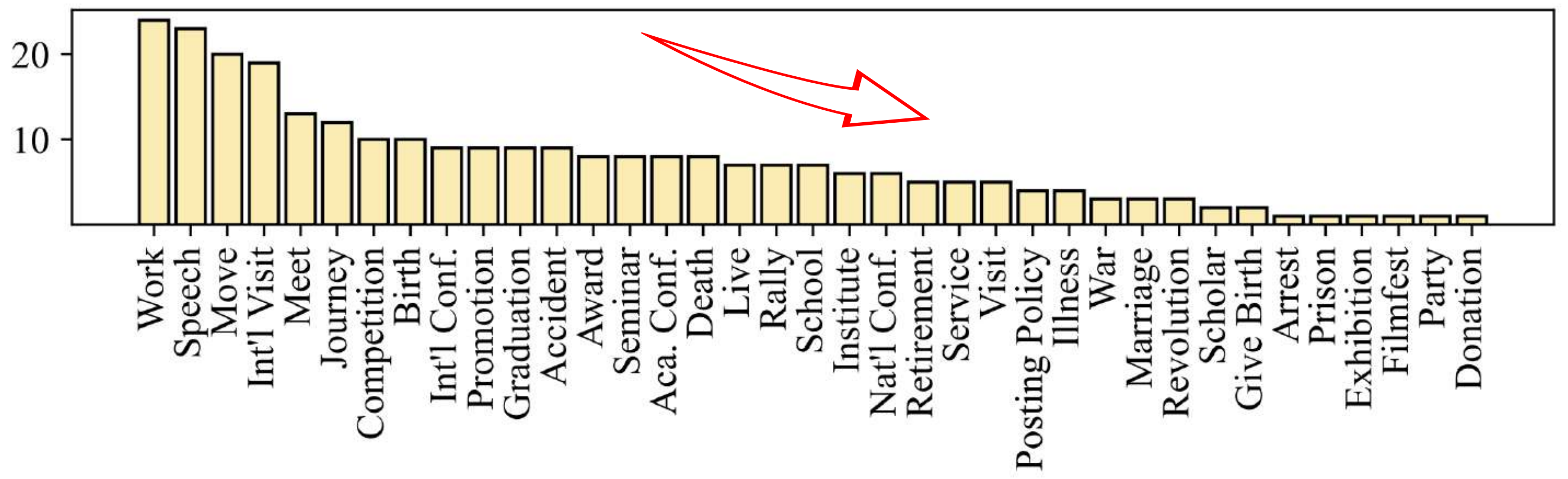
− Get rid of manual rules
− **Specific** population groups
− **Limited** generalization ability

Ours (COSMOS)

✓ Directly classify trajectory triplets
✓ Combine contrastive learning and semi-supervised learning to enhance model's **generalization ability**
✓ Extract millions of trajectory data from English Wikipedia biographies

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?

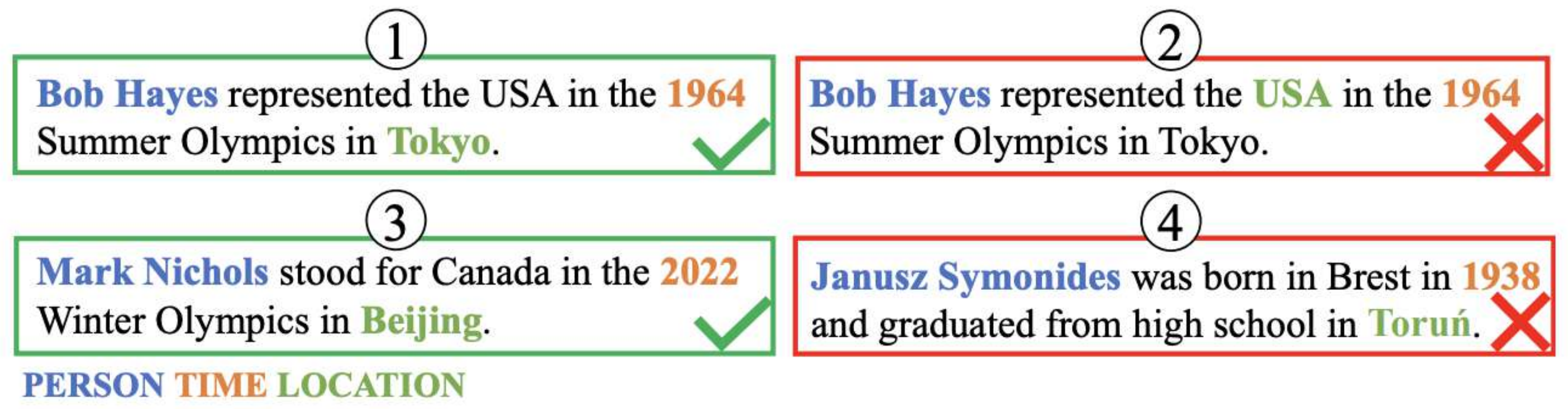


COSMOS

When we delve deeper into the structure between samples…

Similarity
The contexts of snippets (1) and (3) are similar (both about sport events), suggesting the same way of extraction

Dissimilarity
The context of snippet (1) and that of snippet (4) (about birth and study) indicate the way of different extraction pattern

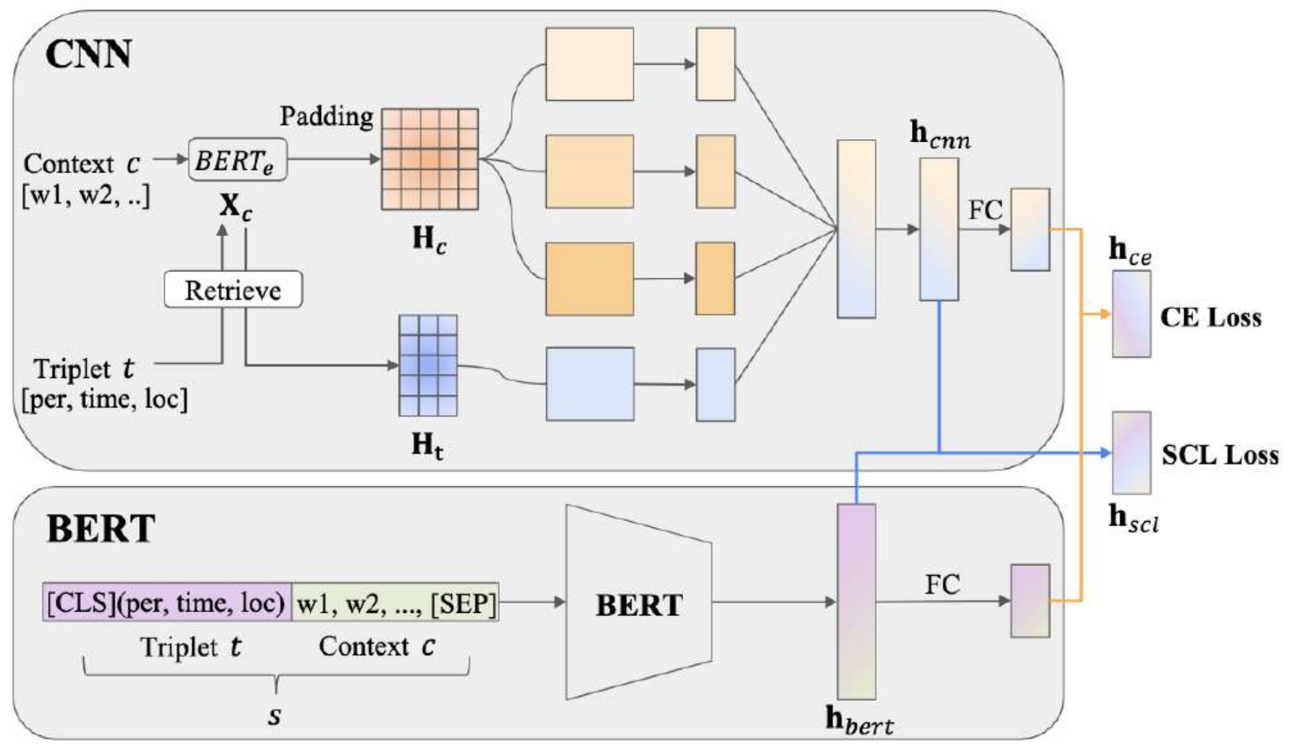

1
Bob Hayes represented the USA in the 1964 Summer Olympics in Tokyo. ✓
2
Bob Hayes represented the USA in the 1964 Summer Olympics in Tokyo. ✗
3
Mark Nichols stood for Canada in the 2022 Winter Olympics in Beijing. ✓
4
Janusz Symonides was born in Brest in 1938 and graduated from high school in Toruń. ✗
PERSON TIME LOCATION

COSMOS (COntrastive learning and Semi-supervised learning MOdel for extracting Spatio-temporal life trajectory)

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

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

- Use **contrastive-learning** to capture intra-sample relation
- Use **semi-supervised learning** to extent the training data



WikiLifeTrajectory Dataset

First, we design a **preprocessing tool** to extract candidate triplets (Person, Time, Location) from biography pages. Our extraction pipeline can cover at least **85%** of the trajectories mentioned on different pages.

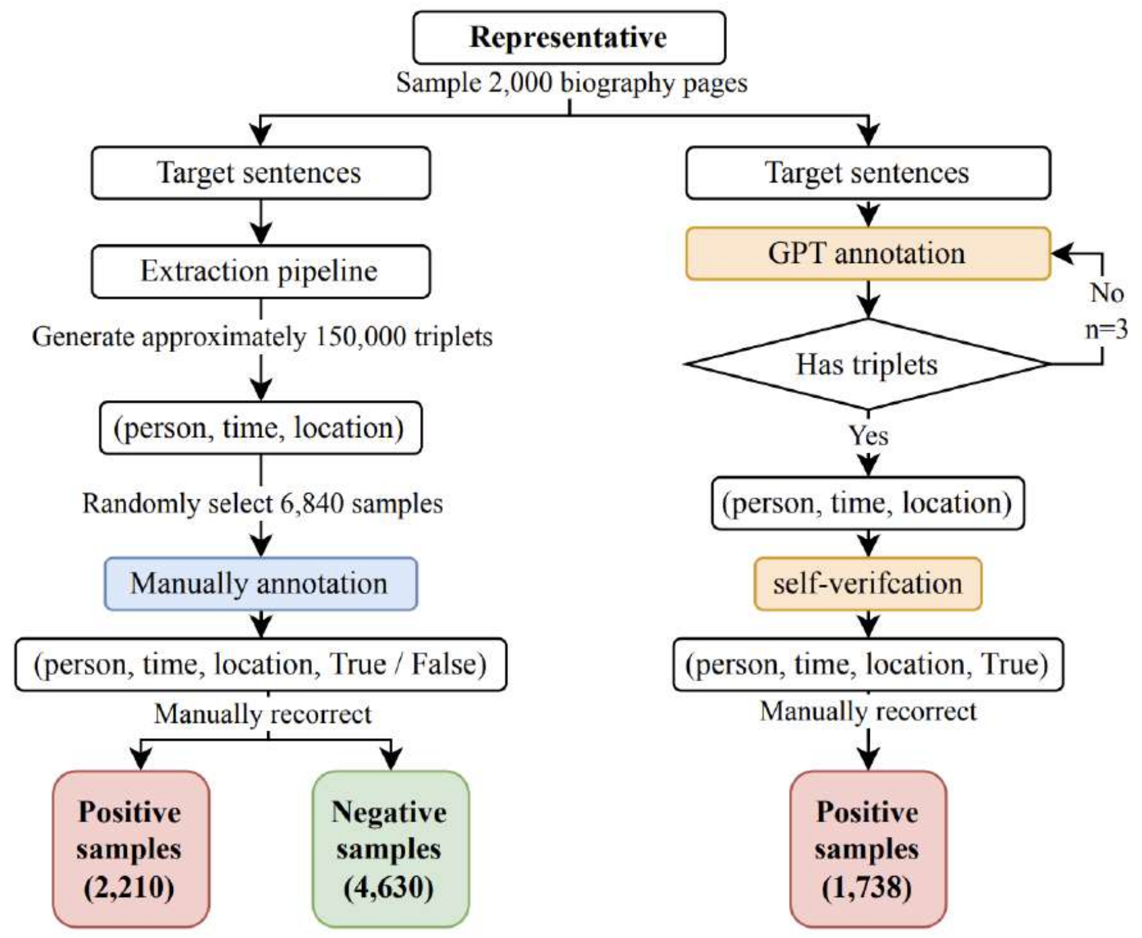


Figure 2: The flowchart illustrates the process of annotating the "Representative" dataset to obtain triplets and their corresponding labels.

Subset 1: Representative

We employ a **stratified sampling** and annotation strategy based on occupation to collect representative trajectories. These samples are labeled by human annotators and GPT-3.5.

Subset 2: Regular

We collect trajectories from another ten biographies (#274) and use them as an independent test set.

Experiments

Prediction & Coverage Performance

	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

Table 1: Performance comparison on the test set. Due to the extreme imbalance between Precision and Recall of GPT-3.5, we specifically highlight the Recall for it with an asterisk (*). Apart from that, the best results are indicated by bold text, while the second-best ones are highlighted with underlines.

Ablation Study

	Representative				Representative$_m$				Representative$_g$	Regular	
	Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)	R (%)	R (%)	Avg-R (std)
COSMOS$_{\text{w/o ssl\&scl}}$	85.23	83.00	85.52	84.24	82.66	71.62	77.89	74.62	**95.52**	68.97	0.6955 ± 0.1791
COSMOS$_{\text{w/o ssl}}$	85.85	85.64	83.33	84.47	83.83	75.33	75.22	75.27	93.96	69.34	0.6636 ± 0.2479
COSMOS$_{\text{w/o scl}}$	86.63	**87.47**	82.91	85.13	**84.80**	**78.07**	74.48	76.23	93.96	71.89	0.6777 ± 0.2109
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

Table 3: Results of the ablation study. Bold text indicates the best results, while underlined text represents the second-best ones.

Analysis of a Sample Set

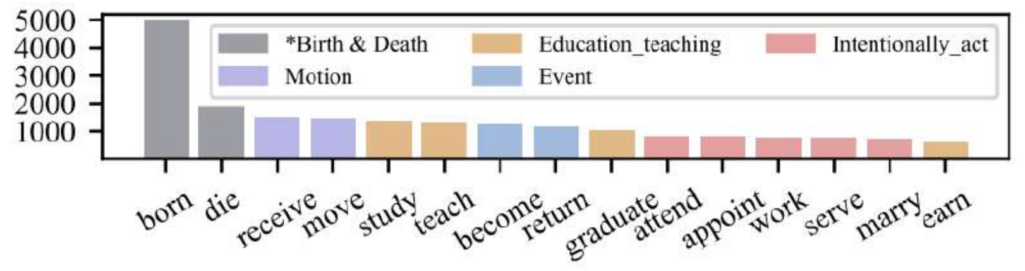


Figure 5: The distribution of the top 15 frequent verbs associated with the trajectories of historians. The horizontal axis represents verbs and the vertical axis represents their corresponding quantities. The * legend indicates the custom category independent of FrameNet.

As a **use case** of our extracted data, we collect 20,786 trajectory triplets for 8,272 historians. We group their **types** and visualize the results at both **individual and group levels**.

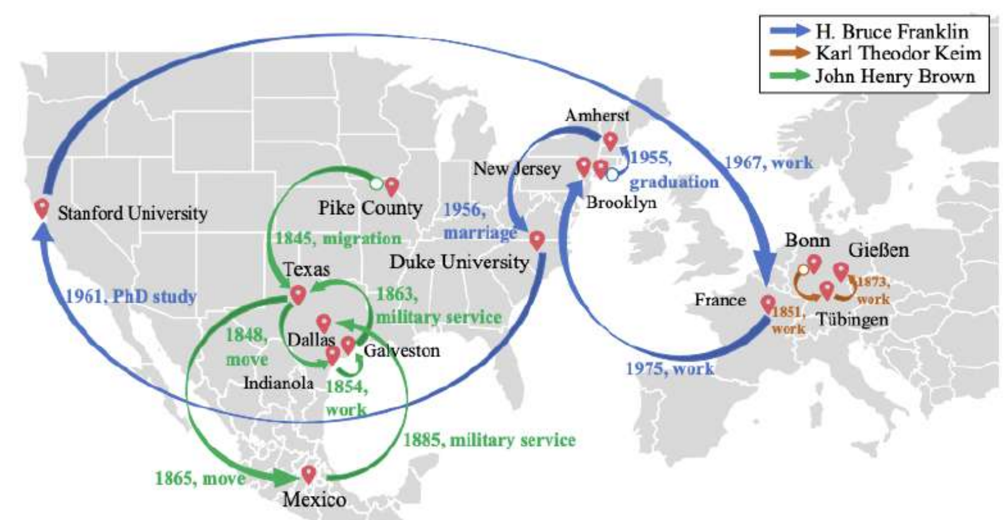


Figure 6: Life trajectories of *H. Bruce Franklin*, *Karl Theodor Keim* and *John Henry Brown*. The arrows of each color represent the life trajectory of the corresponding individual. The start point of each trajectory is marked with a circle. The year and purpose of the move are labeled on the arrows.

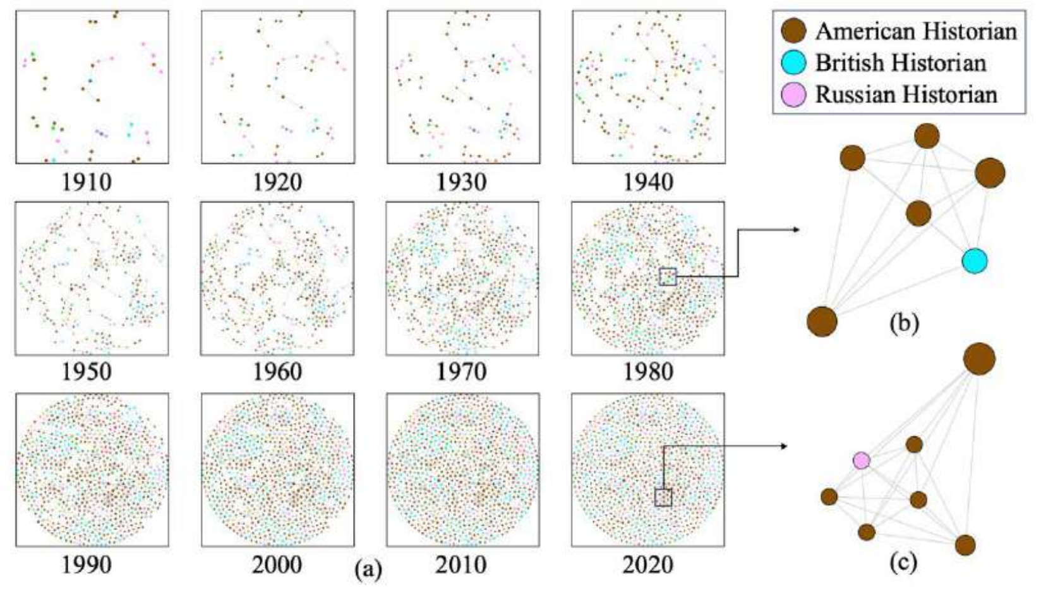


Figure 7: Dynamic interaction network comprising 899 historians. (a) Snapshots of the network every 10 years from 1910 to 2020. Nodes represent historians, the sizes of nodes are the PageRank values, and their nationalities are indicated by colors. The visualization is created using the Fruchterman Reingold layout. (b) and (c) zoom in on two connected components in the 1980 snapshot and 2020 snapshot respectively.

In total, we extract over **five million** trajectories from **1.9 million** Wikipedia biographies — feel free to explore and use the dataset!

Paper Code Dataset

Template Credit: ShanghaiTech-poster-template