CareerMap: Visualizing Career Trajectory

Kan Wu, Jie Tang, Zhou Shao, Xinyi Xu, Bo Gao & Shu Zhao

Dept. Of Computer Science, Tsinghua University

Challenges

Challenge 1:

Name ambiguity

solution → Unified Probabilistic Models [1]

Challenge 2:

Data incompletion

solution → Spatial-Temporal Factor Graph Model (STFGM)

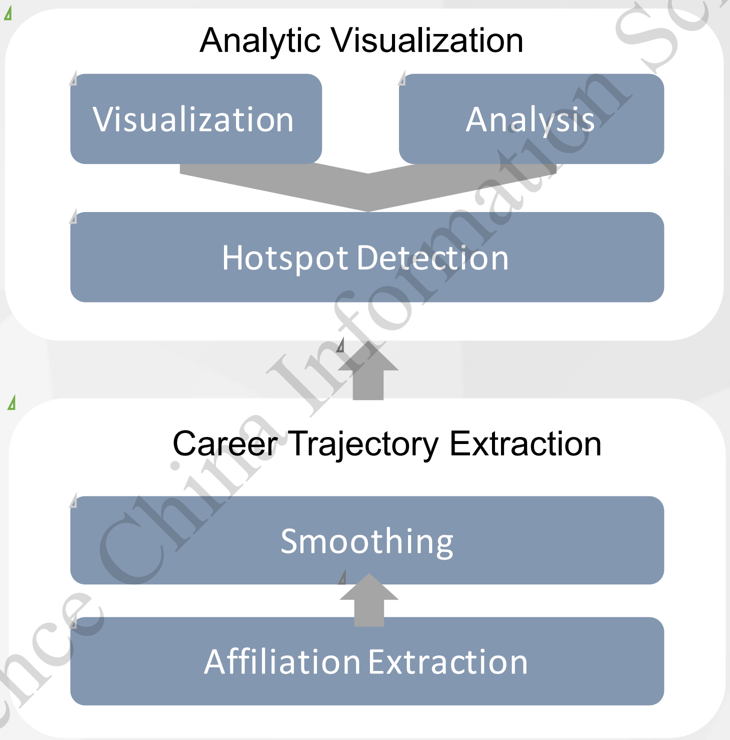
Challenge 3:

Visualize many scholars' merged trajectories on the map, e.g. 100 people move from Boston to NewYork

solution → Hotspot detection algorithm

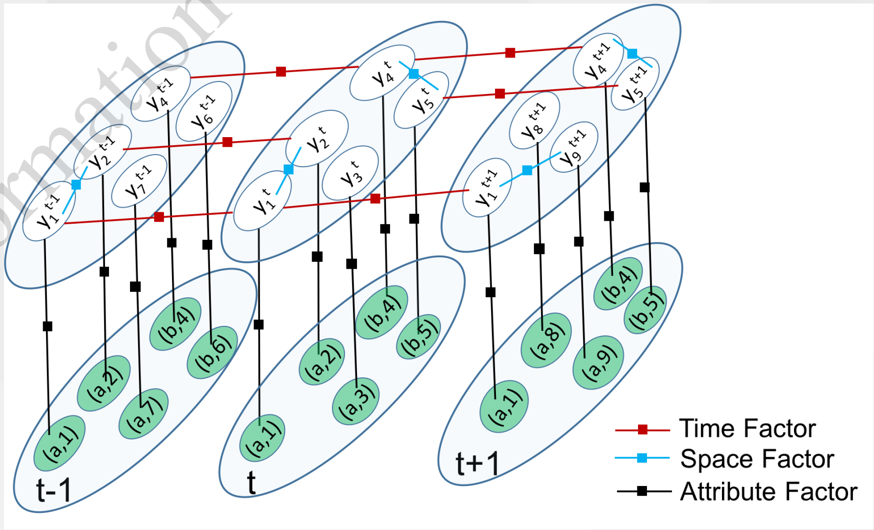
1.Jie Tang, A.C.M. Fong, Bo Wang, and Jing Zhang. A Unified Probabilistic Framework for Name Disambiguation in Digital Library. IEEE TKDE, 2012

Architecture



Spatial-Temporal Factor Graph Model (continued)

- ## The general idea
  - try to find the affiliation-known coauthor who has the same affiliation as the target author with missing affiliation.



Each green point with common t outside, representing a tuple of <Time t, Author a_{i1} , Author ai2 >, is an observation instance where ai1 is the target author and ai2 is a coauthor with known affiliation at t. Associated with each observation instance is a hidden binary-valued variable representing the affiliation similarity between the two authors. If they belong to the same affiliation at that time, the hidden value is 1, otherwise 0.

Spatial-Temporal Factor Graph Model (continued)

- ## Attribute factor
  - captures the features of each tuple <Time t, Author a_{i1} , Author a_{i2}>,

$$f(\boldsymbol{x}_i^t, y_i^t) \triangleq \frac{1}{Z_{\boldsymbol{\omega}}} \exp\left\{\boldsymbol{\omega}^T \boldsymbol{\Phi}(\boldsymbol{x}_i^t, y_i^t)\right\} \quad (1)$$

- ## Space factor
  - captures the correlation between the hidden variables in the same time
  - $\mathcal{N}_S$ denotes all the space relations

$$\mathcal{S}(y_i^t, \mathcal{N}_S(y_i^t)) \triangleq \frac{1}{Z_{\boldsymbol{\beta}}} \exp\left\{\sum_{y_j^t \in \mathcal{N}_S(y_i^t)} \boldsymbol{\beta}^T \boldsymbol{\Psi}(y_i^t, y_j^t)\right\} \quad (2)$$

- ## Time factor
  - captures the correlation between the hidden variables in the same time
  - $\mathcal{N}_T$ denotes all the time relations

$$\mathcal{T}(y_i^t, \mathcal{N}_T(y_i^t)) \triangleq \frac{1}{Z_{\gamma}} \exp\left\{\sum_{y_i^{t'} \in \mathcal{N}_T(y_i^t)} \boldsymbol{\gamma}^T \boldsymbol{\Omega}(y_i^t, y_i^{t'})\right\} \quad (3)$$

Spatial-Temporal Factor Graph Model

• Model Learning

- Maximize the likelihood of the observed data
- $\theta \triangleq (\omega^T, \beta^T, \gamma^T)^T$ is the parameters to be learned of the model

$$P(Y|X,\boldsymbol{\theta}) = \prod_t \prod_i f(\boldsymbol{x}_i^t, y_i^t) \mathcal{S}(y_i^t, \mathcal{N}_S(y_i^t)) \mathcal{T}(y_i^t, \mathcal{N}_T(y_i^t))$$
$$= \frac{1}{Z_\omega Z_\beta Z_\gamma} \exp\left\{ (\boldsymbol{\omega}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}^T) \sum_t \sum_i g(y_i^t) \right\}$$
$$= \frac{1}{Z_\theta} \exp\left\{ \boldsymbol{\theta}^T \mathcal{G}(Y) \right\} \quad (4)$$

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathcal{O}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} logP(Y^L|X,\boldsymbol{\theta}) \quad (5)$$

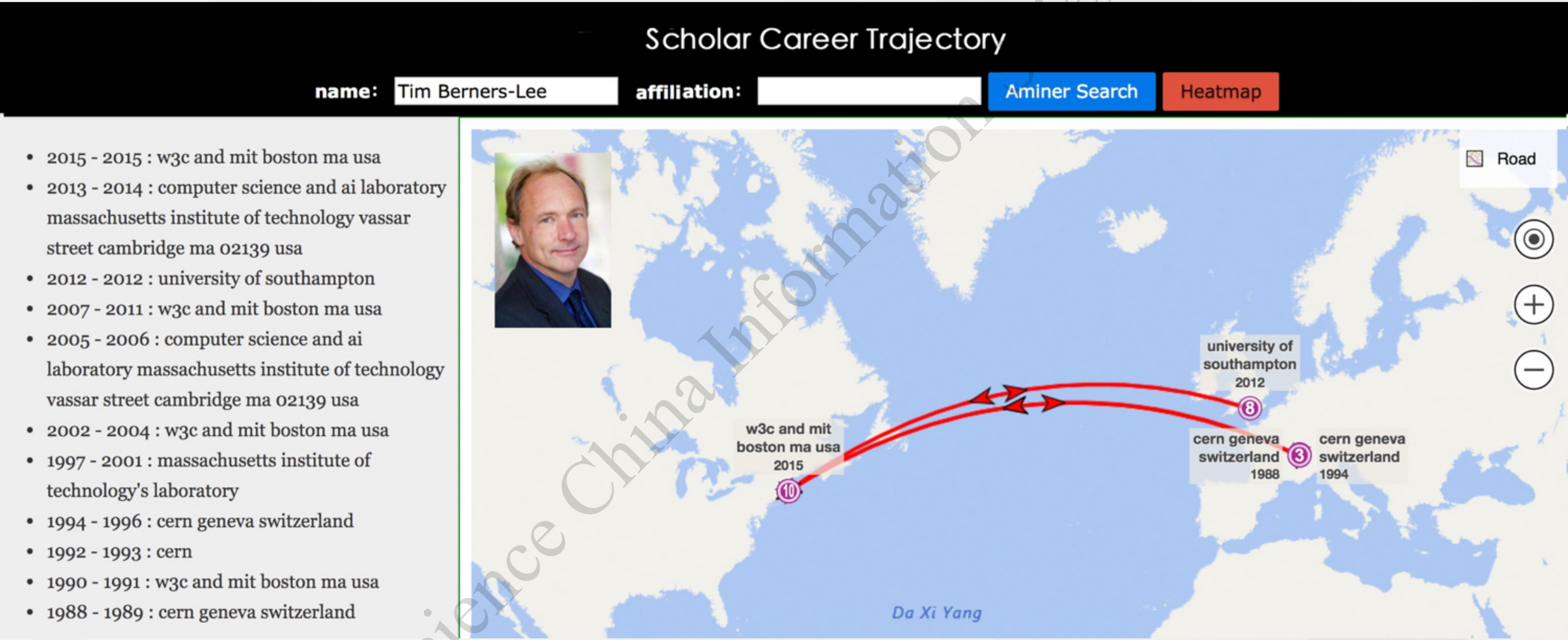
Smooting

- The general idea
  - Use weight to reflect confidence of an affiliation at a time.
  - Leverage the number of papers with the affiliation at time t as the weight.
  - Denoting the weights at t_1 and t_2 are w_1 and w_2 respectively, the weight center t_c can be computed from:

$$\frac{t_c - t_1}{t_2 - t_c} = \frac{w_1}{w_2}$$

  - If information between t_1 and t_2 is missing,
  - $\forall t$ ($t_1 < t < t_c$), Affiliation(a, t) = Affiliation(a, t_1)
  - $\forall t$ ($t_c < t < t_2$), Affiliation(a, t) = Affiliation(a, t_2)

Example of scholar career trajectory extraction



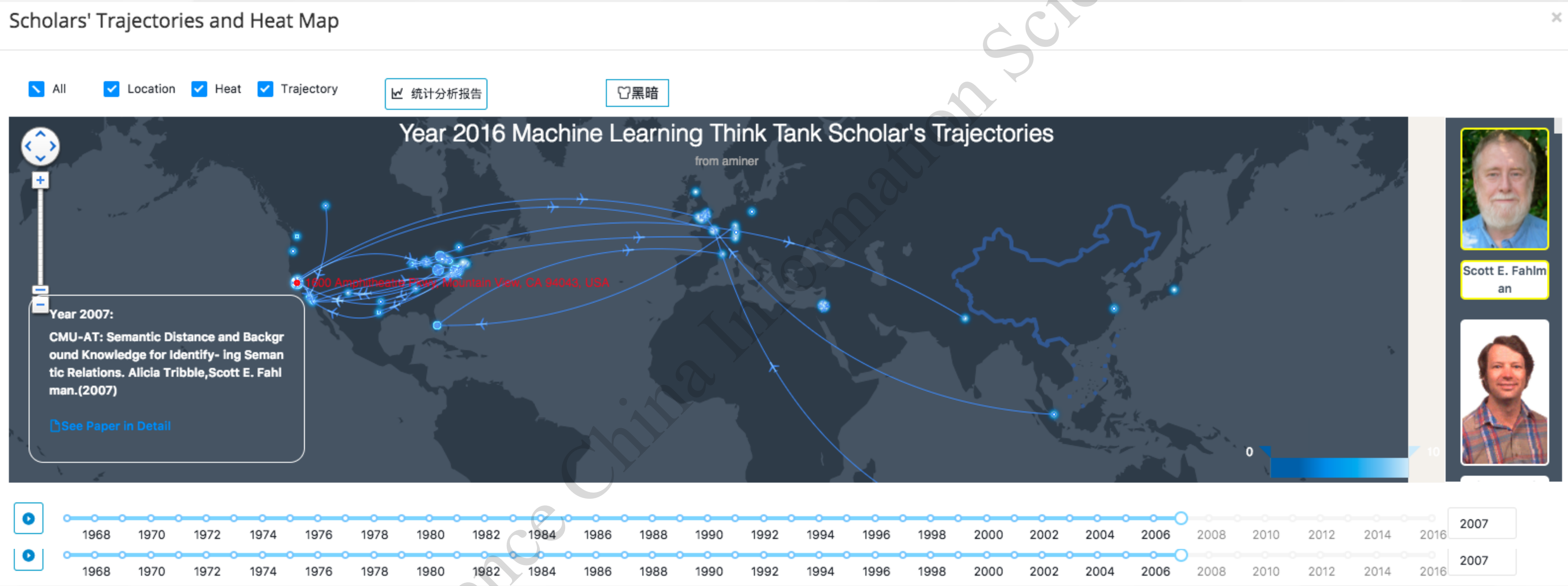
Hotspot detection algorithm

- The general idea
  - The heat centers have more neighbors than surrounding points.
  - The heat centers ”absorb” their surrounding points as their neighbors. If a point is ”absorbed” by a heat center, then its neigh-bors are emptied.
  - Finally, the points left out with nonempty neighbors are heat centers.

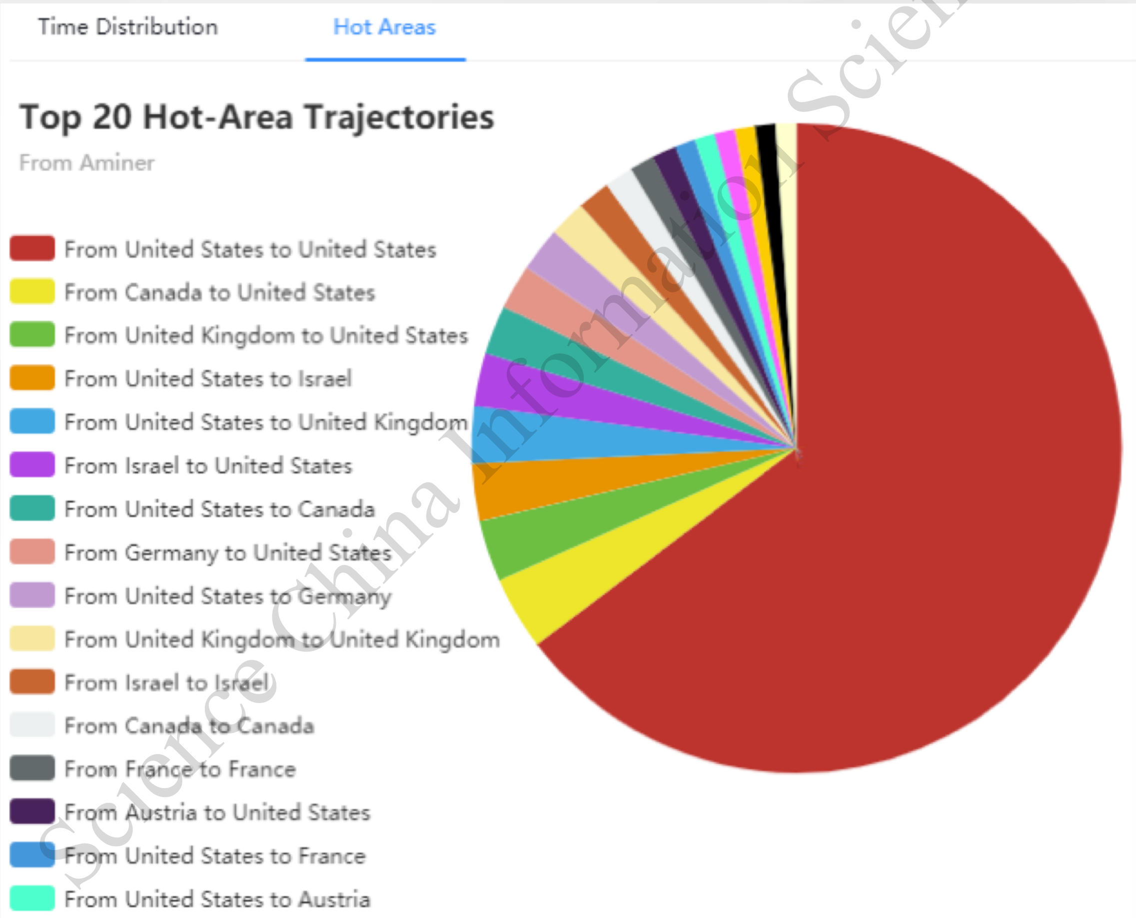
Algorithm 1 Hotspot detection

```
Input  : set of points {v_1, v_2, ..., v_N}, radius of
         heatCenters R
Output: heatCenters and points in them
for i ⇐ 1 to N do
    neighbors(v_i) ⇐ {v_i}
end
for i ⇐ 1 to N − 1 do
    for j ⇐ i + 1 to N do
        if distance(v_i, v_j) < 2R then
            neighbors(v_i) ⇐ neighbors(v_i) ∪ {v_j}
            neighbors(v_j) ⇐ neighbors(v_j) ∪ {v_i}
        end
    end
end
{v'_1, v'_2, ..., v'_N} ⇐ sort {v_1, v_2, ..., v_N} according to
 neighbor size in desc order
for i ⇐ 1 to N − 1 do
    for v_j ∈ neighbors(v'_i) do
        neighbors(v_j) ⇐ ∅
    end
end
for i ⇐ 1 to N do
    if neighbors(v'_i) ≠ ∅ then
        heatCenters(v'_i) = neighbors(v'_i)
    end
end
```

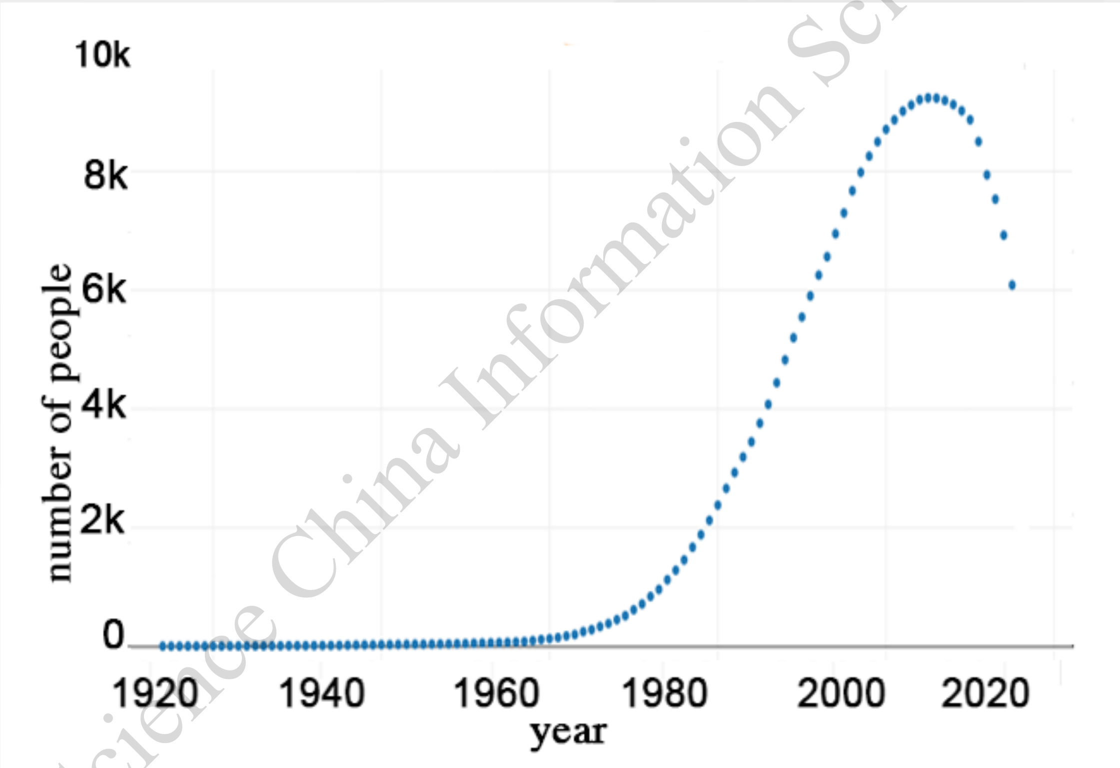
Trajectory map generated by Career Map



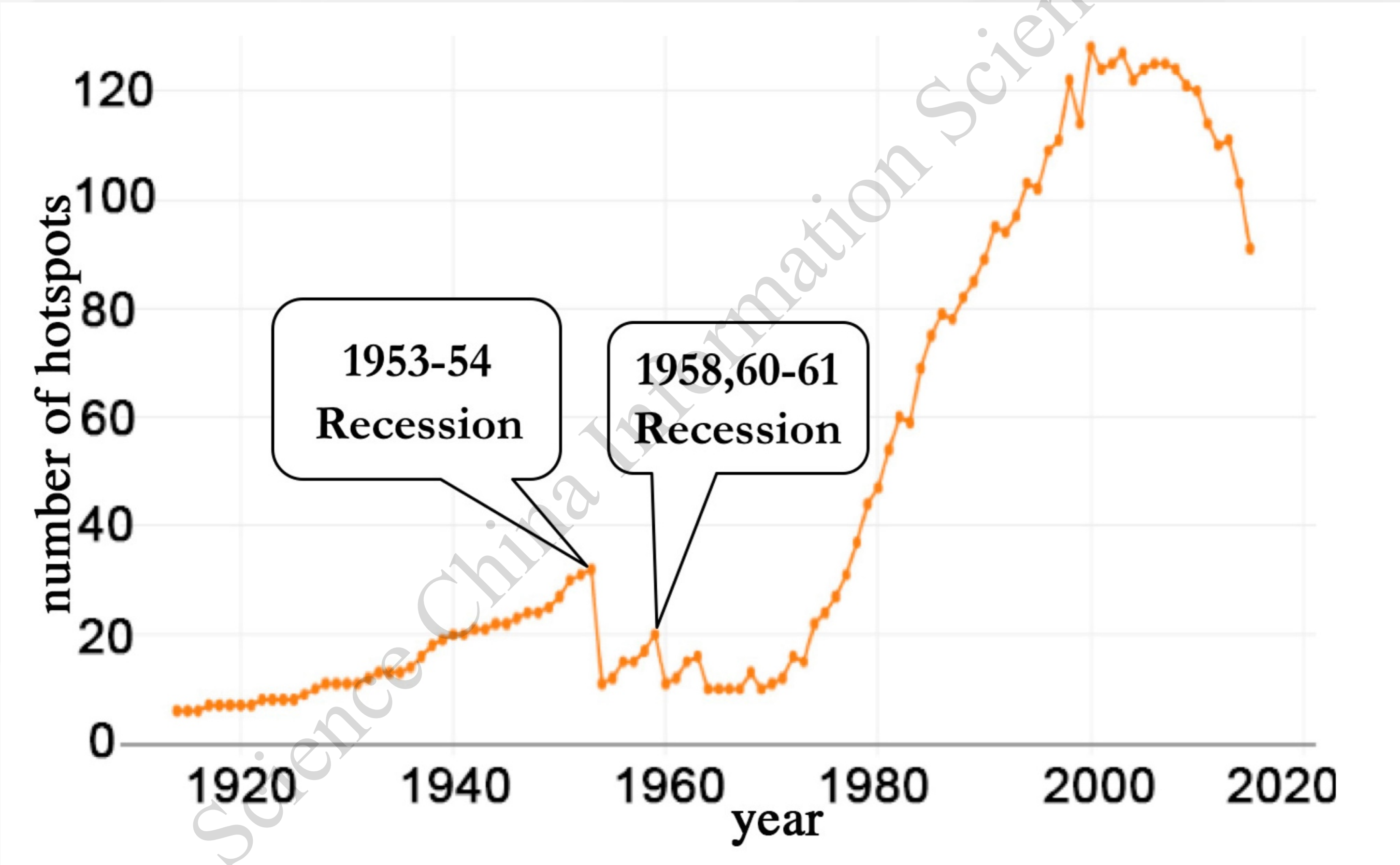
Analytic Visualization



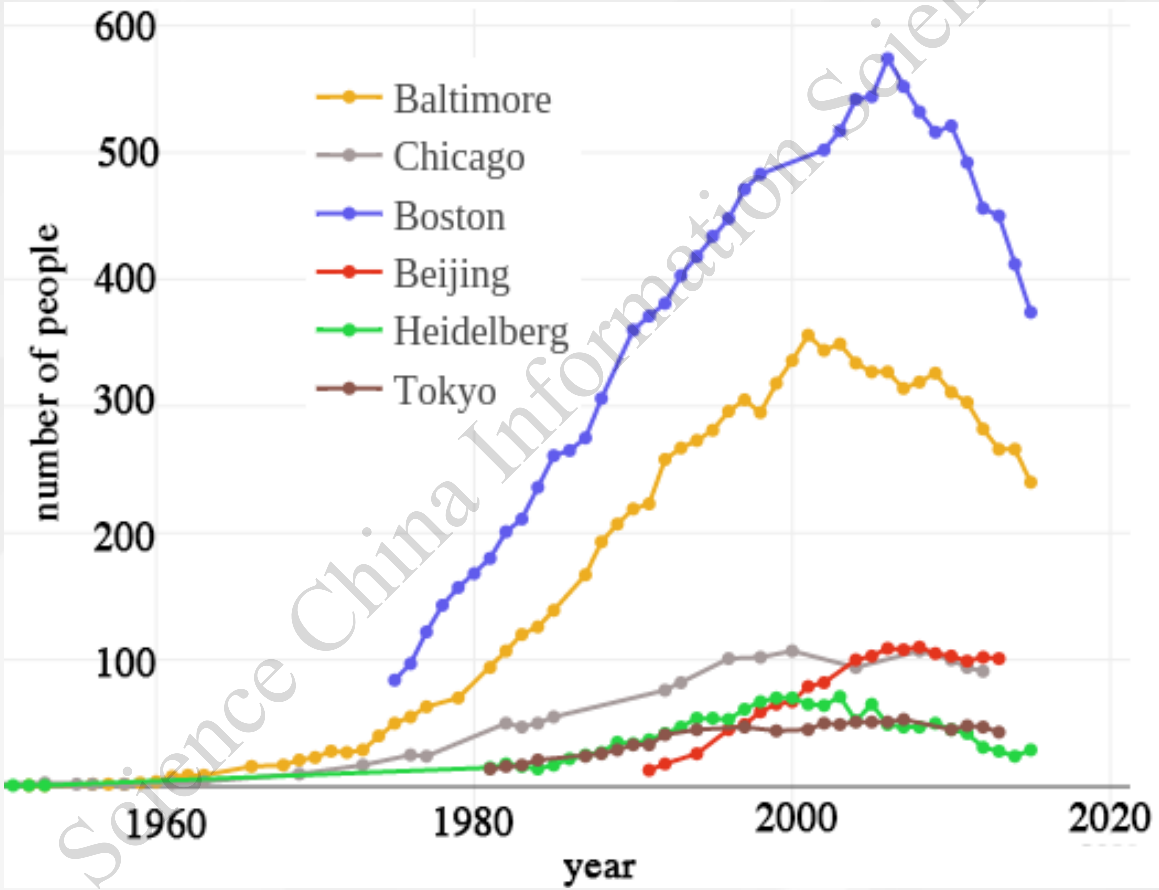
Some Interesting Case Study (continued)



Some Interesting Case Study (continued)



Some Interesting Case Study



Summary

1. We introduce the challenges of building CareerMap, a system for visualizing scholars' career trajectory
2. Architecture ,technologies and main features of the system
3. Some interesting case studies

Thanks

Q&A