CareerMap: Visualizing Career Trajectory

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

Dept. Of Computer Science, Tsinghua University
### Challenges

<table>
<thead>
<tr>
<th>Challenge 1:</th>
<th>Solution</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name ambiguity</td>
<td>Unified Probabilistic Models [1]</td>
<td><strong>Challenges</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Solution</strong></td>
</tr>
<tr>
<td>Challenge 2:</td>
<td>Spatial-Temporal Factor Graph Model (STFGM)</td>
<td><strong>Solution</strong></td>
</tr>
<tr>
<td>Data incompletion</td>
<td></td>
<td><strong>Solution</strong></td>
</tr>
<tr>
<td>Challenge 3:</td>
<td>Hotspot detection algorithm</td>
<td><strong>Solution</strong></td>
</tr>
<tr>
<td>Visualize many scholars’ merged trajectories on the map, e.g. 100 people move from Boston to New York</td>
<td></td>
<td><strong>Solution</strong></td>
</tr>
</tbody>
</table>

Architecture

Analytic Visualization

Visualization

Analysis

Hotspot Detection

Career Trajectory Extraction

Smoothing

Affiliation Extraction
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 \(<\text{Time t, Author } a_{i1}, \text{Author } a_{i2}>\), is an observation instance where \(a_{i1}\) is the target author and \(a_{i2}\) 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.
• **Attribute factor**
  - captures the features of each tuple \(<\text{Time } t, \text{ Author } a_{i1}, \text{ Author } a_{i2}>\),

• **Space factor**
  - captures the correlation between the hidden variables in the same time
  - \(\mathcal{N}_S\) denotes all the space relations

• **Time factor**
  - captures the correlation between the hidden variables in the same time
  - \(\mathcal{N}_T\) denotes all the time relations

---

### Mathematical Formulations

1. **Attribute factor**
   
   \[
   f(x^t_i, y^t_i) \triangleq \frac{1}{Z_\omega} \exp\left\{\omega^T \Phi(x^t_i, y^t_i)\right\}
   \]  
   ![Attribute factor formula]

2. **Space factor**
   
   \[
   S(y^t_i, \mathcal{N}_S(y^t_i)) \triangleq \frac{1}{Z_\beta} \exp\left\{\sum_{y^t_j \in \mathcal{N}_S(y^t_i)} \beta^T \Psi(y^t_i, y^t_j)\right\}
   \]  
   ![Space factor formula]

3. **Time factor**
   
   \[
   T(y^t_i, \mathcal{N}_T(y^t_i)) \triangleq \frac{1}{Z_\gamma} \exp\left\{\sum_{y^t_{i'} \in \mathcal{N}_T(y^t_i)} \gamma^T \Omega(y^t_i, y^t_{i'})\right\}
   \]  
   ![Time factor formula]
Model Learning

- Maximize the likelihood of the observed data
- \( \theta \overset{\Delta}{=} (\omega^T, \beta^T, \gamma^T)^T \) is the parameters to be learned of the model

\[
P(Y|X, \theta) = \prod_t \prod_i f(x_i^t, y_i^t) S(y_i^t, N_S(y_i^t)) T(y_i^t, N_T(y_i^t))
\]

\[
= \frac{1}{Z_\omega Z_\beta Z_\gamma} \exp \left\{ (\omega^T, \beta^T, \gamma^T) \sum_t \sum_i g(y_i^t) \right\}
\]

\[
= \frac{1}{Z_\theta} \exp \left\{ \theta^T g(Y) \right\}
\]  

(4)

\[
\theta^* = \arg \max_{\theta} \mathcal{O}(\theta) = \arg \max_{\theta} \log P(Y^L|X, \theta)
\]  

(5)
• **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,
  - $\forall t (t_1 < t < t_c), \text{Affiliation}(a, t) = \text{Affiliation}(a, t_1)$
  - $\forall t (t_c < t < t_2), \text{Affiliation}(a, t) = \text{Affiliation}(a, t_2)$
Hotspot detection algorithm

- 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 neighbors 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
  \( \text{neighbors}(v_i) = \{v_i\} \)
end

for \( i = 1 \) to \( N - 1 \) do
  for \( j = i + 1 \) to \( N \) do
    if distance\((v_i, v_j) < 2R\) then
      \( \text{neighbors}(v_i) = \text{neighbors}(v_i) \cup \{v_j\} \)
      \( \text{neighbors}(v_j) = \text{neighbors}(v_j) \cup \{v_i\} \)
    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 \( v_j \in \text{neighbors}(v_i') \) do
    \( \text{neighbors}(v_j) \leftarrow \emptyset \)
  end
end

for \( i = 1 \) to \( N \) do
  if \( \text{neighbors}(v_i') \neq \emptyset \) then
    \( \text{heatCenters}(v_i') = \text{neighbors}(v_i') \)
  end
end
Trajectory map generated by Career Map
Top 20 Hot-Area Trajectories

- From United States to United States
- From Canada to United States
- From United Kingdom to United States
- From United States to Israel
- From United States to United Kingdom
- From Israel to United States
- From United States to Canada
- From Germany to United States
- From United States to Germany
- From United Kingdom to United Kingdom
- From Israel to Israel
- From Canada to Canada
- From France to France
- From Austria to United States
- From United States to France
- From United States to France
- From United States to Austria
Some Interesting Case Study (continued)
Some Interesting Case Study (continued)

[Graph showing the number of hotspots over years with notable recesions in 1953-54 and 1958-60-61.]
Some Interesting Case Study
We introduce the challenges of building CareerMap, a system for visualizing scholars’ career trajectory.

Architecture, technologies and main features of the system

Some interesting case studies