Disambiguation by Namesake Risk Assessment

Thorsten Doherr | Leibniz Centre for European Economic Research (ZEW), Germany
Symposium on Entity Resolution
26th March 2021 | USPTO
US7855906B2
DC bus voltage control for two stage solar converter

Classifications: H02J 3/04
H02M 7/537
H02M 7/539

Inventors: Anthony Michael Klodowski
David Smith
Robert Gregory Wagoner
Sidney Allen Barker

Applicant: General Electric Company

US7239036B2
System and method for power control in wind turbines

Classifications: F03D 9/00
H02P 9/04

Inventors: John Douglas D'Atre
Anthony Michael Klodowski
Allen Michael Ritter
David Smith
Robert Gregory Wagoner
Luis Jose Garces
Henning Luetze

Applicant: General Electric Company
The risk of erroneously linking two documents of namesakes (homonymous individuals), depends on...

- the rarity/commonness of the name → number of namesakes in the population
- and the number of individuals sharing the same properties defining the linkage → peer group.

We need to find a way to predict the number of namesakes for any name!

We need to find a way to assess the risk of linking documents of namesakes!

We need to find a way to calculate the peer group size for any intersection of document properties!
Likelihood of Not Drawing a Namesake

$$1 - \left( \frac{N - n}{N - 1} \right)$$

1. \(\frac{N - n}{N - 1} = \frac{100 - 5}{100 - 1} = \frac{95}{99}\)

2. \(\frac{N - n - 1}{N - 2} = \frac{94}{98}\)

3. \(\frac{N - n - 2}{N - 3} = \frac{93}{97}\)

$$\frac{N - n - s + 2}{N - s + 1} = \frac{87}{91}$$

$$\frac{95 \cdot 94 \cdot 93 \cdot 92 \cdot 91 \cdot 90 \cdot 89 \cdot 88 \cdot 87}{99 \cdot 98 \cdot 97 \cdot 96 \cdot 95 \cdot 94 \cdot 93 \cdot 92 \cdot 91} \approx 0.68$$

\(N: \) population, e.g. 100  
\(n: \) namesakes in population, e.g. 5  
\(s: \) peer group size, e.g. 10
\[ P_{risk}(n, s) = 1 - P_{unique}(n, s) \]

\[ = 1 - \prod_{i=0}^{s-2} \frac{N - n - i}{N - 1 - i} \]

\[ = 1 - \frac{(N - n)!/(N - n - s + 1)!}{(N - 1)!/(N - s)!} \]

\[ = 1 - \exp\left(\ln f(N - n) - \ln f(N - n - s + 1) - \ln f(N - 1) + \ln f(N - s)\right) \]

Stirling’s approximation for \( \ln(x!) \):

\[ \ln f(x) = \left(x + \frac{1}{2}\right) \ln(x + 1) - (x + 1) + \frac{1}{2} \ln(2\pi) \]
### Namesake Indicator: Percentile Rank of Minimum Occurrence

<table>
<thead>
<tr>
<th>Aggregated Name List</th>
<th>Count Firstname</th>
<th>Count Lastname</th>
<th>Dense Rank</th>
<th>minocc</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Firstname</strong></td>
<td><strong>Lastname</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JON</td>
<td>OPSAL</td>
<td>1,062</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>ZACHARY B</td>
<td>SMITH</td>
<td>13</td>
<td>6,739</td>
<td>13</td>
</tr>
<tr>
<td>HIROTO</td>
<td>YOSHII</td>
<td>311</td>
<td>184</td>
<td>184</td>
</tr>
<tr>
<td>ANTONY</td>
<td>BUTLER</td>
<td>2,647</td>
<td>851</td>
<td>787</td>
</tr>
<tr>
<td>ADAM</td>
<td>MILLER</td>
<td>1,302</td>
<td>3,973</td>
<td>1,011</td>
</tr>
<tr>
<td>DAVID</td>
<td>SMITH</td>
<td>11,746</td>
<td>6,739</td>
<td>1,334</td>
</tr>
</tbody>
</table>

---

ZEW
Namesake Prediction

1. Aggregation of a master sample of unambiguously identified individuals on the name level counting namesakes to determine minocc.

<table>
<thead>
<tr>
<th>German Stakeholder Database</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>6,731,543</td>
</tr>
<tr>
<td>Names</td>
<td>4,691,779</td>
</tr>
<tr>
<td>Max. namesakes per name</td>
<td>1,118</td>
</tr>
<tr>
<td>Avg. namesakes per name</td>
<td>1.43</td>
</tr>
</tbody>
</table>

2. Weighted Poisson regression of namesakes on a polynomial of minocc:

$$namesakes = \exp(\beta_0 + \beta_1 minocc + \beta_2 minocc^2 + \beta_3 minocc^3 + \cdots + \beta_5 minocc^5) + \varepsilon$$

[frequency weight: namesakes]
Namesake Prediction Result

- prediction
- 95% confidence band
- extrapolation (1000 namesakes)
- lower bound

ZEW
## Aggregated Name List

<table>
<thead>
<tr>
<th>Firstname</th>
<th>Lastname</th>
<th>Count Firstname</th>
<th>Count Lastname</th>
<th>Dense Rank</th>
<th>minocc</th>
<th>n</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>JON</td>
<td>OPSAL</td>
<td>1,062</td>
<td>3</td>
<td>3</td>
<td>0.00223</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ZACHARY B</td>
<td>SMITH</td>
<td>13</td>
<td>6,739</td>
<td>13</td>
<td>0.00966</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HIROTO</td>
<td>YOSHII</td>
<td>311</td>
<td>184</td>
<td>184</td>
<td>0.13670</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>ANTONY</td>
<td>BUTLER</td>
<td>2,647</td>
<td>851</td>
<td>787</td>
<td>0.58470</td>
<td>16</td>
<td>42</td>
</tr>
<tr>
<td>ADAM</td>
<td>MILLER</td>
<td>1,302</td>
<td>3,973</td>
<td>1,011</td>
<td>0.75111</td>
<td>34</td>
<td>92</td>
</tr>
<tr>
<td>DAVID</td>
<td>SMITH</td>
<td>11,746</td>
<td>6,739</td>
<td>1,334</td>
<td>0.99109</td>
<td>313</td>
<td>865</td>
</tr>
</tbody>
</table>
Peer Group Intersection

**IPC: H02**

Patents: 123944  
Names: 124622  
Peers: 127622  
\( P_{risk}(313, 127622) = 99.7\% \)

**General Electric Company**

Patents: 21364  
Names: 17148  
Peers: 17200  
\( P_{risk}(313, 17200) = 55.0\% \)

**Intersection**

Patents: 1928  
Names: 2116  
Peers: 2167  
\( P_{risk}(313, 2167) = 9.6\% \)
Structural Complexity

Entity Relationship Diagram
Patstat – Autumn 2020
Mutual Traits

US7855906B2

US7239036B2
Mutual Trait Combinations of a Namespace
Disambiguation by Namespace Traversal
## Rapid Deployment

### Office EPO DPMA USPTO JPO

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Patents</td>
<td>3,101,170</td>
<td>3,512,309</td>
<td>5,282,235</td>
<td>10,625,369</td>
</tr>
<tr>
<td>Inventor records</td>
<td>8,002,041</td>
<td>6,912,139</td>
<td>12,178,282</td>
<td>22,940,716</td>
</tr>
<tr>
<td>Names</td>
<td>2,196,294</td>
<td>1,866,808</td>
<td>2,603,181</td>
<td>1,963,483</td>
</tr>
<tr>
<td>Individuals</td>
<td>2,681,181</td>
<td>2,324,510</td>
<td>3,299,199</td>
<td>3,954,470</td>
</tr>
<tr>
<td>Name density</td>
<td>1.22</td>
<td>1.25</td>
<td>1.27</td>
<td>2.01</td>
</tr>
<tr>
<td>Mobility*</td>
<td>50,844</td>
<td>22,433</td>
<td>67,051</td>
<td>low data quality</td>
</tr>
<tr>
<td>Threshold</td>
<td>2.5%[10%</td>
<td>(\Delta t \leq 3y)]</td>
<td>2.5%[10%</td>
<td>(\Delta t \leq 3y)]</td>
</tr>
<tr>
<td>Min namesakes</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

*An individual is considered mobile if the career spans at least two countries.*
Pros & Cons

- **No training data required!**
  - No tedious manual disambiguation
  - No inherent namesake bias

- **Rapid deployment on any person related database:**
  - Patent data
  - Publication data
  - Career linkage, e.g. inventors to authors to founders based on aggregated career data

- **Transparency: every link can be deconstructed**
  - No shrugging about incomprehensible ML decisions
  - No “Clever Hans” shenanigans: obtuse shortcuts taken by ML approaches

- **Hyper-parameters, like lower bound, extrapolation and threshold, would benefit from unbiased training data.**

- **Clustering of soft properties is much more involved than distance calculations, all the more if topics, e.g. of abstracts, have to be clustered.**

- **Relative distance relations, e.g. time gap between patents, do not have a relatable peer group. Workarounds are based on overlapping grids and lack the refinement.**
Thank you for your attention
Time for questions

Disambiguation by Namesake Risk Assessment
https://www.zew.de/publikationen/disambiguation-by-namesake-risk-assessment

Clustering Tool
https://github.com/ThorstenDoherr/searchengine
Magnitude of the Namesake Bias

![Graph showing the share of names and the share of namesake afflicted individuals against sample size. The y-axis represents share ranging from 0 to 1, and the x-axis represents sample size ranging from 100 to 6,731,543.]
Distance Based Relations

Overlapping Time Windows

Overlapping Spatial Grids

key trait
1 TIMA20002002
1 TIMB19972002
1 TIMB20002005
2 TIMA20032005
2 TIMB20002005
2 TIMB20032008
3 TIMA20062008
3 TIMB20032008
3 TIMB20062011

key trait
1 GRDA0202
1 GRDB0101
1 GRDB0201
1 GRDB0102
1 GRDB0202
2 GRDA0302
2 GRDB0201
2 GRDB0301
2 GRDB0202
2 GRDB0302