Distance decay in anonymous Wikipedia authorship

Darren Hardy
Ph.D. candidate, Bren School, UCSB

23 February 2010, ThinkSpatial

Advisor: Prof. Frew (Bren)
VGI RA: Profs. Goodchild, Elwood (UW), and Sui (OSU).
Agenda

• Geotagged Wikipedia and collective authorship

• Does distance matter to online authors?

• Study with data over 7 years and 21 languages

• Methods and metrics

• Results
Collective authorship of place

• Rise of a global digital commons of geographic knowledge

• Volunteered geographic information from distributed contributors

• *Collective authorship* is a “mass collective effort by individuals to produce information artifacts within a digital commons.” (Hardy 2008)
Wikipedia

- Wikipedia.org, an online collaborative encyclopedia since 2001

- **Popular**...
  Ranked #6 by popularity. 42% of external traffic via Google.
  In 2009, 365 million unique visitors, 133.6 billion page views

- **Vast**...
  15.0 million articles in 272 languages

- **Anyone can edit**...
  860 million edits by 22.2 million contributors
  1,076,908 “Wikipedians” (10+ edits), and
  91,817 “active” Wikipedians (5+ edits per month)

Wikipedia authorship

- **Registered** authors
  - Only username required
  - Name, email, etc. optional
  - IP address kept hidden

- **Anonymous** authors
  - IP address made public
  - But nothing else

- **Administrative** authors
  - Privileged registered user

- **Bot** authors
  - Automated program
Wikipedia authorship

- **Registered** authors
  - Only username required
  - Name, email, etc. optional
  - IP address kept hidden

- **Anonymous** authors
  - IP address made public
  - But nothing else

- **Administrative** authors
  - Privileged registered user

- **Bot** authors
  - Automated program

### Contributions to “Copenhagen Opera House”

<table>
<thead>
<tr>
<th># of Contributions</th>
<th>Username or IP</th>
<th>Most Recent</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>Dybdahl</td>
<td>18-Sep-2005</td>
</tr>
<tr>
<td>6</td>
<td>85.233.237.71 (anon)</td>
<td>12-Jan-2008</td>
</tr>
<tr>
<td>3</td>
<td>Viva-Verdi</td>
<td>8-Sep-2006</td>
</tr>
<tr>
<td>1</td>
<td>Hemmingsen</td>
<td>3-Jan-2007</td>
</tr>
<tr>
<td>4</td>
<td>81.62.92.47 (anon)</td>
<td>15-Apr-2006</td>
</tr>
<tr>
<td>1</td>
<td>Thue</td>
<td>28-Feb-2006</td>
</tr>
<tr>
<td>2</td>
<td>Ghent</td>
<td>30-Apr-2006</td>
</tr>
<tr>
<td>3</td>
<td>Valentinian</td>
<td>7-Jan-2007</td>
</tr>
<tr>
<td>3</td>
<td>83.77.92.205 (anon)</td>
<td>10-Apr-2006</td>
</tr>
<tr>
<td>3</td>
<td>130.226.234.229 (anon)</td>
<td>29-Sep-2007</td>
</tr>
<tr>
<td>2</td>
<td>86.149.109.196 (anon)</td>
<td>15-Oct-2007</td>
</tr>
<tr>
<td>2</td>
<td>Uppland</td>
<td>24-Dec-2005</td>
</tr>
<tr>
<td>2</td>
<td>87.48.100.222 (anon)</td>
<td>12-Jan-2006</td>
</tr>
</tbody>
</table>
What is geotagging?

- Marks location on Earth

- **coordinates**
  "55.7° N, 12.6° E"

- maybe, a place name
  "Copenhagen, Denmark"

- maybe, a type
  "City"

You are here

Bartholl 2006
Geotagged Wikipedia

- WikiProject: Geographic coordinates “WP:Geo”
  - Provides templates for tagging geographic coordinates to articles
  - Thus, a subset of Wikipedia is “geotagged”
- Example: Wiki markup

```
{{Infobox...
|latd = 34 |latm = 25 |lats = 33 |latNS = N
|longd = 119 |longm = 42 |longs = 51 |longEW = W
...}}

''Santa Barbara'' is a city in [[Santa Barbara County, California]], [[United States]]. Situated on an east-west trending section of coastline...
The Copenhagen Opera House (in Danish usually called Operahuset) is the national opera house of Denmark, and among the most modern opera houses in the world. It is also one of the most expensive opera houses ever built with construction costs well over 500 million U.S. dollars[1]. It is located on the island of Holmen in the center of Copenhagen.
Copenhagen Opera House

The Copenhagen Opera House (in Danish usually called Operaen) is the national opera house of Denmark, and among the most modern opera houses in the world. It is also one of the most expensive opera houses ever built with construction costs well over 500 million U.S. dollars[^1]. It is located on the island of Holmen in the center of Copenhagen.

![The Copenhagen Opera House](image)

[^1]: http://example.com/costs

**Contents**

1. History
2. Location
3. Characteristics
4. Takkelloftet
5. Backstage
6. Architecture, art and decoration
7. Transport
8. References
9. External links

**History**
<table>
<thead>
<tr>
<th>Global systems</th>
<th>Denmark</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Service</strong></td>
<td><strong>Service</strong></td>
</tr>
<tr>
<td>ACME Mapper</td>
<td>Findvej.dk</td>
</tr>
<tr>
<td>ArcWeb Explorer</td>
<td>Findvej.dk with Wikipedia Map</td>
</tr>
<tr>
<td>Bing Maps</td>
<td>Eniro</td>
</tr>
<tr>
<td>Blue Marble Navigator</td>
<td>Satellite</td>
</tr>
<tr>
<td>ExploreOurPlan.net</td>
<td>Daily</td>
</tr>
<tr>
<td>Flash Earth</td>
<td>Satellite</td>
</tr>
<tr>
<td>Fourmilab</td>
<td>Satellite</td>
</tr>
<tr>
<td>GeoBros</td>
<td>Satellite</td>
</tr>
<tr>
<td>GeoNames</td>
<td>Satellite</td>
</tr>
<tr>
<td>GlobeXplorer</td>
<td>Satellite</td>
</tr>
<tr>
<td>Google Earth</td>
<td>Open</td>
</tr>
<tr>
<td>Google Maps</td>
<td>Map</td>
</tr>
<tr>
<td>GPS Visualizer</td>
<td>Map</td>
</tr>
<tr>
<td>Map24</td>
<td>Map</td>
</tr>
<tr>
<td>MapQuest</td>
<td>Map</td>
</tr>
<tr>
<td>MapTech</td>
<td>Map</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Map</strong></td>
<td><strong>Map</strong></td>
</tr>
<tr>
<td><strong>Satellite</strong></td>
<td><strong>Satellite</strong></td>
</tr>
<tr>
<td><strong>Terrain</strong></td>
<td><strong>Terrain</strong></td>
</tr>
<tr>
<td><strong>Mapnik</strong></td>
<td><strong>Mapnik</strong></td>
</tr>
<tr>
<td><strong>Flash required</strong></td>
<td><strong>Flash required</strong></td>
</tr>
<tr>
<td><strong>Bird’s Eye</strong></td>
<td><strong>Bird’s Eye</strong></td>
</tr>
</tbody>
</table>

View the location above by selecting a mapping provider.
GeoHack - Copenhagen Opera House

WGS84  55° 40' 55.1 N, 12° 36' 2.9 E
55.681944, 12.600556

UTM  33U 489141 6173291

Zoom  8  Scale  ±1:10000
Region  DK-101  Type  landmark
Title  Copenhagen Opera House

Contents  Global and Local systems - Wikipedia articles - Photos - Other - Export

View the location above by selecting a mapping provider:

Global systems

<table>
<thead>
<tr>
<th>Service</th>
<th>Map</th>
<th>Satellite</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACME Mapper</td>
<td>Map</td>
<td>Satellite</td>
<td>Terrain, Mapsik</td>
</tr>
<tr>
<td>ArcWeb Explorer</td>
<td>Map</td>
<td>Satellite</td>
<td>(Flash required)</td>
</tr>
<tr>
<td>Bing Maps</td>
<td>Map</td>
<td>Aerial</td>
<td>Bird's Eye</td>
</tr>
<tr>
<td>Blue Marble Navigator</td>
<td></td>
<td>Satellite</td>
<td></td>
</tr>
<tr>
<td>ExploreOurPla.net</td>
<td></td>
<td>Daily</td>
<td></td>
</tr>
<tr>
<td>Flash Earth</td>
<td></td>
<td>Satellite</td>
<td></td>
</tr>
<tr>
<td>Fourmilab</td>
<td></td>
<td>Satellite</td>
<td></td>
</tr>
<tr>
<td>GeoBios</td>
<td></td>
<td>Satellite</td>
<td></td>
</tr>
<tr>
<td>GeoNames</td>
<td></td>
<td>Satellite</td>
<td></td>
</tr>
<tr>
<td>Globexplorer</td>
<td></td>
<td>Satellite</td>
<td></td>
</tr>
<tr>
<td>Google Earth</td>
<td></td>
<td>Open</td>
<td>w/ meta data</td>
</tr>
<tr>
<td>Google Maps</td>
<td></td>
<td>Satellite</td>
<td>Terrain</td>
</tr>
<tr>
<td>GPS Visualizer</td>
<td></td>
<td>Satellite</td>
<td>Terrain</td>
</tr>
<tr>
<td>Map24</td>
<td></td>
<td>Map</td>
<td></td>
</tr>
<tr>
<td>MapQuest</td>
<td></td>
<td>Map</td>
<td></td>
</tr>
<tr>
<td>MapTech</td>
<td></td>
<td>Map</td>
<td></td>
</tr>
</tbody>
</table>

Denmark

<table>
<thead>
<tr>
<th>Service</th>
<th>Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Findvej.dk</td>
<td>Map</td>
</tr>
<tr>
<td>Findvej.dk with Wikipedia</td>
<td>Map</td>
</tr>
<tr>
<td>Eniro</td>
<td>Map</td>
</tr>
</tbody>
</table>

Indirect

- Kort & Maritkelstyrelsen Topo
- Krak.dk Map

View all other regional systems
Google Earth articles embedded in map
Google Earth articles embedded in map
Scale of geotagged Wikipedia

# of contributions, authors, and articles (log scale)

growth = power law
Wikipedia Growth Animated


(Zachte, 2009)
Does distance matter in Wikipedia authorship?
Distance matters

• The First Law of Geography: “Everything is related to everything else, but near things are more related than distant things.” (Tobler, 1970, p. 236)

• Many phenomena have distance decay functions

  • Power distance decay model
e.g., transportation

  \[ I_{ij} = \alpha d_{ij}^{-\beta} \]

  • Exponential distance decay model
e.g., migration, diffusion

  \[ I_{ij} = \alpha e^{-\beta d_{ij}} \]

  • Combined, “gamma” model

  \[ I_{ij} = \alpha d_{ij}^{-\beta} e^{-\gamma d_{ij}} \]
Does distance matter to online authors?

• Does TFL apply to Wikipedia authorship? TFL predicts authors should write about nearby places more than distant places. If so, what is the distance decay function?

• Hypothesis -- distance still matters in online knowledge creation

  • H1: Anonymous Wikipedia authors write more about nearby places
    \[ Pr(d) > Pr(d + \delta d) \]

  • H2: and follows an exponential distance decay function:
    \[ Pr(d) = \alpha e^{-\beta d} \]
Signature distance metric

San Francisco: 446 km
Santa Barbara
San Diego: 307 km
Signature distance metric

1 article

San Francisco
446 km

Santa Barbara

San Diego
307 km
Signature distance metric

1 article

2 authors

San Francisco
446 km

Santa Barbara

San Diego
307 km
Signature distance of 1 article with 2 authors

San Francisco
446 km

Santa Barbara

San Diego
307 km
Signature distance of 1 article with 2 authors

San Francisco 446 km
Santa Barbara
San Diego 307 km

$$\bar{d} = \frac{446 + 307}{2} = 377 \text{ km}$$
Signature distance of 1 article with 2 authors

San Francisco
446 km

Santa Barbara

San Diego
307 km

2 contributions

5 contributions

\[
\bar{d}_{signature} = \frac{2 \cdot 446 + 5 \cdot 307}{2 + 5} = 347 \text{ km}
\]
Signature distance metric for article $\alpha$

$$D(\alpha) = \sum_i (w(\alpha, \rho_i) \cdot d(\alpha, \rho_i))$$

signature distance
weighted average distance
Signature distance metric for article $\alpha$

$$D(\alpha) = \sum_i (w(\alpha, \rho_i) \cdot d(\alpha, \rho_i))$$

- Signature distance
- Weighted average distance

Weight is percentage of work:
$$w(\alpha, \rho_i) = \frac{n(\alpha, \rho_i)}{\sum_i n(\alpha, \rho_i)}$$

Distance:
$$d(\alpha, \rho_i) = \text{GREATCIRCLEDISTANCE}(\alpha, \rho_i)$$
Oat Mountain
2 anonymous authors with 2 revisions; signature distance = 54 km
University of California, Santa Barbara
135 anonymous authors with 719 revisions; signature distance = 533 km
University of California, Santa Barbara (German)
10 anonymous authors with 18 revisions; signature distance = 7,988 km
Tibet Autonomous Region
114 anonymous authors with 210 revisions; signature distance = 8,980 km
How do authors geotag articles?

- **Manual**
  - Author uses online mapping software or other method to get coordinates
  - Then, inserts the coordinates into an article:
    - Mount Everest is at {{coord|27|59|16|N|86|56|40|E}}

- **Automated**
  - en:User:The Anomebot2 cross-references articles with gazetteers
  - Maybe-Checker searches for non-geotagged articles
How do we locate anonymous authors?

• **IP Address Geolocation**

  • Convert IP to (lat, lon) using *GeoLite City database* from MaxMind, Inc.
    - Free version of commercial GeoIP product
    - Proprietary technology, but methods include “user-entered location data.”
    - GeoLite seeded with public data; GeoIP seeded with proprietary data

  • Accuracy: % of IP addresses resolved within 25 miles of true location
    - US (79%), Germany (71%), France (60%), Australia (59%),
      Japan (54%), and UK (54%).

  • 2.6 million IP addresses converted into 45k geographic coordinates.
Example GeoIP lookup

<table>
<thead>
<tr>
<th>Hostname</th>
<th>Country Code</th>
<th>Country Name</th>
<th>Region Name</th>
<th>City</th>
<th>Postal Code</th>
<th>Latitude Longitude</th>
<th>ISP</th>
<th>Organization</th>
<th>Metro Code</th>
<th>Area Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>128.111.110.49 US</td>
<td>United States</td>
<td>CA</td>
<td>California</td>
<td>Santa Barbara</td>
<td>93106</td>
<td>34.4329 -119.8372</td>
<td>University of California, Santa Barbara</td>
<td>University of California, Santa Barbara</td>
<td>855</td>
<td>805</td>
</tr>
</tbody>
</table>

These results were generated with the [Perl API](https://perldoc.perl.org) and the commercial [GeoIP City](https://www.maxmind.com), [GeoIP ISP](https://www.maxmind.com), and [GeoIP Organization](https://www.maxmind.com) databases.
Data collection and sampling
Data collection

authors & readers

Wikipedia

wikipedia.org

DB

MySQL (Florida, USA)
Data collection

authors & readers

Wikipedia

wikipedia.org DB
MySQL (Florida, USA)

replication

toolserver.org DB
MySQL (Germany)
100s of DBs; per language
Data collection

authors & readers

Wikipedia

wikipedia.org DB

MySQL (Florida, USA)

Extraction

MySQL (Santa Barbara)

articles

authors

geotags

SQL

toolserver.org DB

replication

MySQL (Germany)

100s of DBs; per language
Data collection

authors & readers

Wikipedia

wikipedia.org DB

MySQL (Florida, USA)

Extraction

signatures

MySQL (Santa Barbara)

articles

authors

geotags

SQL

toolserver.org DB

MySQL (Germany)

100s of DBs; per language

MySQL (Florida, USA)

replication
Data collection

authors & readers

Wikipedia

wikipedia.org DB

MySQL (Florida, USA)

My research

Extraction

signatures

MySQL (Santa Barbara)

articles

authors

geotags

toolserver.org DB

MySQL (Germany)
100s of DBs; per language

replication

SQL
Data in study

Authors

Excluded 581,530

Included 2,845,054

Restrict to anonymous, for location estimates
Data in study

Authors

- Excluded: 581,530
- Included: 2,845,054

Restrict to anonymous, for location estimates

Articles

- Excluded: 438,078
- Included: 550,444

Restrict to articles with anonymous authorship and a single geotag
Excluded 581,530
Included 2,845,054

Restrict to anonymous, for location estimates

Excluded 550,444
Included 438,078

Restrict to articles with anonymous authorship and a single geotag

Excluded 24,810,458
Included 7,285,137

Contributions
Articles with geotags

# of articles per unit area (log scale, 0.1° resolution)

988,522 articles
103,291 distinct locations
Articles in study
require single geotag and anonymous contributions

438,077 articles
85,389 distinct locations

# of articles per unit area (log scale, 0.1° resolution)
Authors in study require anonymous contributions to articles in study

7,285,137 contributions
2,845,054 IP addresses
30,376 distinct locations
World population

# of people per square km (log scale, 2.5’ resolution)

Source: GPWv3 (CIESIN 2005)
Results

Signature distance and simulation
Distribution of signature distances
(n = 438,077 articles)
64% of articles at 2,000 km or less
64% of articles at 2,000 km or less

\[ Pr(D(\alpha) = d) = e^{-\beta D(\alpha)} \]
64% of articles at 2,000 km or less

exponential decay

\[ Pr(D(\alpha) = d) = e^{-\beta D(\alpha)} \]
Distribution of signature distances over 2,000 km
(n = 159,016 articles; 36%)

Distance between Europe and North America
Distribution of article signature distances for 21 languages

Exponential fit at $R^2 = 0.90$
English data

$Pr(D=d) \log_{10} \text{scale}$

$d, \text{distance (km)}$
Japanese data
Exponential distance decay per language

<table>
<thead>
<tr>
<th>Language</th>
<th>Fit ($R^2$)</th>
<th># of contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalan</td>
<td>0.78</td>
<td>286,902</td>
</tr>
<tr>
<td>Chinese</td>
<td>0.76</td>
<td>426,099</td>
</tr>
<tr>
<td>Czech</td>
<td>0.79</td>
<td>327,976</td>
</tr>
<tr>
<td>Danish</td>
<td>0.83</td>
<td>188,464</td>
</tr>
<tr>
<td>Dutch</td>
<td>0.82</td>
<td>1,893,004</td>
</tr>
<tr>
<td><strong>English</strong></td>
<td><strong>0.94</strong></td>
<td><strong>10,727,409</strong></td>
</tr>
<tr>
<td>Esperanto</td>
<td>0.84</td>
<td>377,663</td>
</tr>
<tr>
<td>Finnish</td>
<td>0.85</td>
<td>384,377</td>
</tr>
<tr>
<td>French</td>
<td>0.89</td>
<td>3,196,281</td>
</tr>
<tr>
<td>German</td>
<td>0.86</td>
<td>5,093,034</td>
</tr>
<tr>
<td>Icelandic</td>
<td>0.67</td>
<td>52,016</td>
</tr>
<tr>
<td>Italian</td>
<td>0.84</td>
<td>2,005,277</td>
</tr>
<tr>
<td><strong>Japanese</strong></td>
<td><strong>0.62</strong></td>
<td><strong>916,866</strong></td>
</tr>
<tr>
<td>Norwegian</td>
<td>0.86</td>
<td>389,075</td>
</tr>
<tr>
<td>Polish</td>
<td>0.84</td>
<td>1,356,350</td>
</tr>
<tr>
<td>Portuguese</td>
<td>0.83</td>
<td>1,204,770</td>
</tr>
<tr>
<td>Russian</td>
<td>0.84</td>
<td>726,064</td>
</tr>
<tr>
<td>Slovak</td>
<td>0.68</td>
<td>182,694</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.86</td>
<td>1,680,446</td>
</tr>
<tr>
<td>Swedish</td>
<td>0.84</td>
<td>480,817</td>
</tr>
<tr>
<td>Turkish</td>
<td>0.75</td>
<td>200,011</td>
</tr>
</tbody>
</table>
Exponential distance decay per language
What if distance didn’t matter?

- Simulation

- Replace each author with a randomly selected author

- Recompute article signature distance
Distribution of signature distances
(n = 438,077 articles)

$P(x)$, probability that an article has signature distance $x$

$x$, signature distance (kilometers)
64% of articles at 2,000 km or less but only 10% in simulation.
Summary

• Geospatial analysis of Wikipedia authorship
  • Data over 7 years and 1m geotagged articles in 21 languages
  • Methods:
    • IP geolocation to estimate location of 2.8 million anonymous authors
    • Signature distance metric

• Evidence suggests distance matters in online authorship
  • Results find geographic effects in anonymous authorship
  • Follows exponential distance decay
  • Non-geographic simulation underestimates nearby contributions
Questions?

Email: dhardy@bren.ucsb.edu

Web: www.bren.ucsb.edu/~dhardy

Demo: toolserver.org/~drh08
Thanks.