Introducing the Author-ity Exporter, and a case study of geo-temporal movement of authors

Mikko S. Tuomela, Brent D. Fegley
Illinois Informatics Institute
School of Information Sciences
University of Illinois at Urbana-Champaign

Vetle I. Torvik
School of Information Sciences
University of Illinois at Urbana-Champaign
vtorvik@illinois.edu

ABSTRACT

We introduce a web service, Author-ity Exporter, that permits searching and exporting data from Author-ity -- a database that has PubMed author names disambiguated with a high degree of accuracy [1]. Each author is represented by a cluster of papers annotated by publication count, time-span, affiliations, topics, journals, co-authors, citations as well as imputed data from MapAffil [2], Genni [3], and Ethnea [4] and links to their NIH/NSF grants and USPTO patents; and we have plans for more. This service should enable and simplify new types of author-centered bibliometric analyses with a unique strength in funding, geography, and diversity (gender, ethnicity, and professional age). We also present an illustrative case study of modeling of authors' career movements to and from a specific city based on data retrieved from Author-ity Exporter. The service (and the R code used in the case study) are available at http://abel.ischool.illinois.edu/cgi-bin/exporter/search.pl.

INTRODUCTION

The Author-ity Exporter can be used to access information on authors in the PubMed database up to July 2009 (the current limit of Author-ity; but a newer version is forthcoming). It connects authors in PubMed with information about associated patents, grants, and papers; and it offers an efficient way to produce datasets for further analysis.

Author-ity Exporter accommodates searching PubMed either directly (via "PubMed Search", where a simple text box allows entry and full use of the special PubMed query syntax[1]) or indirectly (via "Search", where text boxes above fields designated for tabular output may be used to constrain or otherwise filter query results). The latter is illustrated in Fig. 1, where the search is for authors who have at least 25 publications and who have "Champaign" among their top 20 affiliation words across all their papers (a rough approximation for having been affiliated with the city of Champaign at some point in their careers).

Fields in Author-ity presently include number of publications, author name, first and last year of publication, affiliations words, topics (Medical Subject Headings), co-authors as well as grants, patents and papers associated with an author. Search results are presently limited to 5,000 authors for technical/practical reasons; nevertheless, all results may be saved for future recall in the "search history" with an option to declare a username.

More information about authors than is displayed in the web interface can be extracted using one of four export buttons. Each export button produces a different data set that can be downloaded for further processing and analysis:

- **Export people**: One row per author; contains the information in the search results, additionally predicted gender and ethnicity using Genni and Ethnea [3, 4].
- **Export papers**: Information on each publication from each author in the search results, including the journal, volume, and issue, as well as affiliation, city, and latitude/longitude of the city as identified by MapAffil [2].
- **Export grants**: For each author, this export lists all funding granted by the US NIH and NSF to the author/principal investigator, including funding amount and affiliation.
- **Export patents**: For each author, this export lists all patents granted by the USTPO to the author/inventor, and more detailed author information.

All export functions produce a tab-limited text file that can be read by a spreadsheet program, such as Microsoft Excel, or an analytical tool, such as R. The four datasets are linked by Author-ity ID, which uniquely identifies each author. It consists of the PubMed ID of their first publication and their position in that publication, e.g, "1234567_2" means the second author on PubMed ID 1234567.

SIG/MET Workshop at ASIS&T, Oct. 14, 2016, Copenhagen, Denmark.

[1] http://www.ncbi.nlm.nih.gov/books/NBK3827/

THE CASE STUDY

Our goal here is to characterize the geographic movement of scientists, when and how far they travelled, to and from Urbana-Champaign, IL. Using Author-ity Exporter, we searched for authors affiliated with Champaign at any point in their careers, and with at least 25 papers to ensure a reasonable number of data points for analysis. For this set of authors, all their papers were downloaded by clicking the "Export papers" button (Fig. 1). The resulting file contains information on 49,044 papers by 785 distinct authors.

A PROBABILISTIC MODEL OF GEOGRAPHIC MOVEMENTS

An author's movements through cities over time are modelled via multiclass logistic regression. Each city is modeled as a sigmoid function of time with two inflections, capturing movements into and out of the city (Equation 1).

$$logit(\Pr\{city_i\}) \sim \beta_0 + \beta_1 year + \beta_2 year^2, \quad (1)$$
$$city_i \in cities$$

Here, $\Pr\{city_i\}$ denotes the probability of $city_i$ and is a function of *year*. The $year^2$ term introduces the second inflection. The parameters β_0, β_1, and β_2 are estimated in a one-versus-rest framework and then probabilities across all *cities* in a given year are normalized so they add up to 1.

The model was implemented as an R script that permits calculating each author's probability of residing in a certain city in a certain year and predicting the city for each year in their careers. Fig. 2 shows the fitted distributions for a specific author. It also permits identifying instances where the author moved to or from a given city and can calculate distances between origin and destination because the exported data has latitudes and longitudes of each city. The mean distance of moves *to Urbana-Champaign* is 2,685 km, while moves *from Urbana-Champaign* is slightly smaller, 2,452 km. The distribution of distances moved is shown in Fig. 3.

It should be noted that the exported free-form text of affiliations have been pre-processed by MapAffil [2], however, because PubMed affiliations are often lacking for pre-1988 or for non-first authors[2], they have been supplemented with affiliations from PubMed Central, NIH grants, and Microsoft Academic Graph, denoted by the prefixes FROMPMC, FROMNIH, FROMPAT, respectively. The R script permits weighting these sources, and PMC is currently down-weighted because it tends to be less reliable in assigning affiliations to a specific author on a paper.

EVALUATION

There are many potential ways of measuring the accuracy and usefulness of the probabilistic model. How often is the model correct (and the bibliographic record wrong), and how often does the model provide a city prediction when none is present? How often are the city-pair movements correctly identified, and how accurate is the predicted year of a move? We are in the process of performing systematic evaluations along those lines. What follows is a preliminary analysis.

Fig. 2 shows the predicted cities over time for an arbitrarily selected author, Douglas Lauffenburger (Author-ity ID = 395370_1). Among 246 papers published during 1979-2009, 92 have affiliation information. The model predicted that he moved from Philadelphia, PA to Urbana-Champaign, IL in 1991, and then to Cambridge, MA in 1996. This aligns well with his public CV except that he moved to UIUC in 1990 (not 1991) and MIT in 1995 (not 1996). These "permanent" moves were correctly identified but predicted a year late, presumably because of the normal time-lag from when the work was performed until publication. The model correctly filtered out minor probability spikes for Minneapolis, MN in 1990 and Boston, MA in 2008 because other cities dominated in those time periods, and it imputed the city for the 154 papers that lacked affiliation information. In a random sample of 50 papers taken from the entire Champaign dataset, we found 4 cases where the model prediction differed from the listed city of affiliation. This rate is similar to that of the author analyzed above where the model differed in 2 out of 92 cases.

In summary, the sigmoids-based probabilistic model accurately captures the geographic movements of authors. It can also predict an author's city of affiliation when it is missing or inaccurately recorded in the bibliography. This should improve spatial scientometrics [5] broadly and enable large-scale studies of geographic mobility in particular.

ACKNOWLEDGEMENTS

NIH P01AG039347. We also thank numerous colleagues on the project who graciously performed initial testing and provided feedback on the Author-ity Exporter.

REFERENCES

[1] Torvik, VI, Smalheiser NR. Author name disambiguation in MEDLINE. ACM TKDD 2009; 3(3), 11.

[2] Torvik VI. MapAffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide. D-Lib Magazine 2015; 21(11/12).

[3] Smith BN, Singh M, Torvik VI. A search engine approach to estimating temporal changes in gender orientation of first names. ACM/IEEE Joint Conf. on Digital Libraries July 22-26, 2013; Indianapolis, IN, USA; 199-208.

[4] Torvik VI, Agarwal S. Ethnea -- an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database. International Symposium on Science of Science March 22-23, 2016; Library of Congress, Washington DC, USA. http://hdl.handle.net/2142/88927

[5] Frenken K, Hardeman S, Hoekman J. Spatial scientometrics: Towards a cumulative research program. Journal of Informetrics 2009; 3(3), 222-232.

[2] https://www.nlm.nih.gov/bsd/mms/medlineelements.html#ad

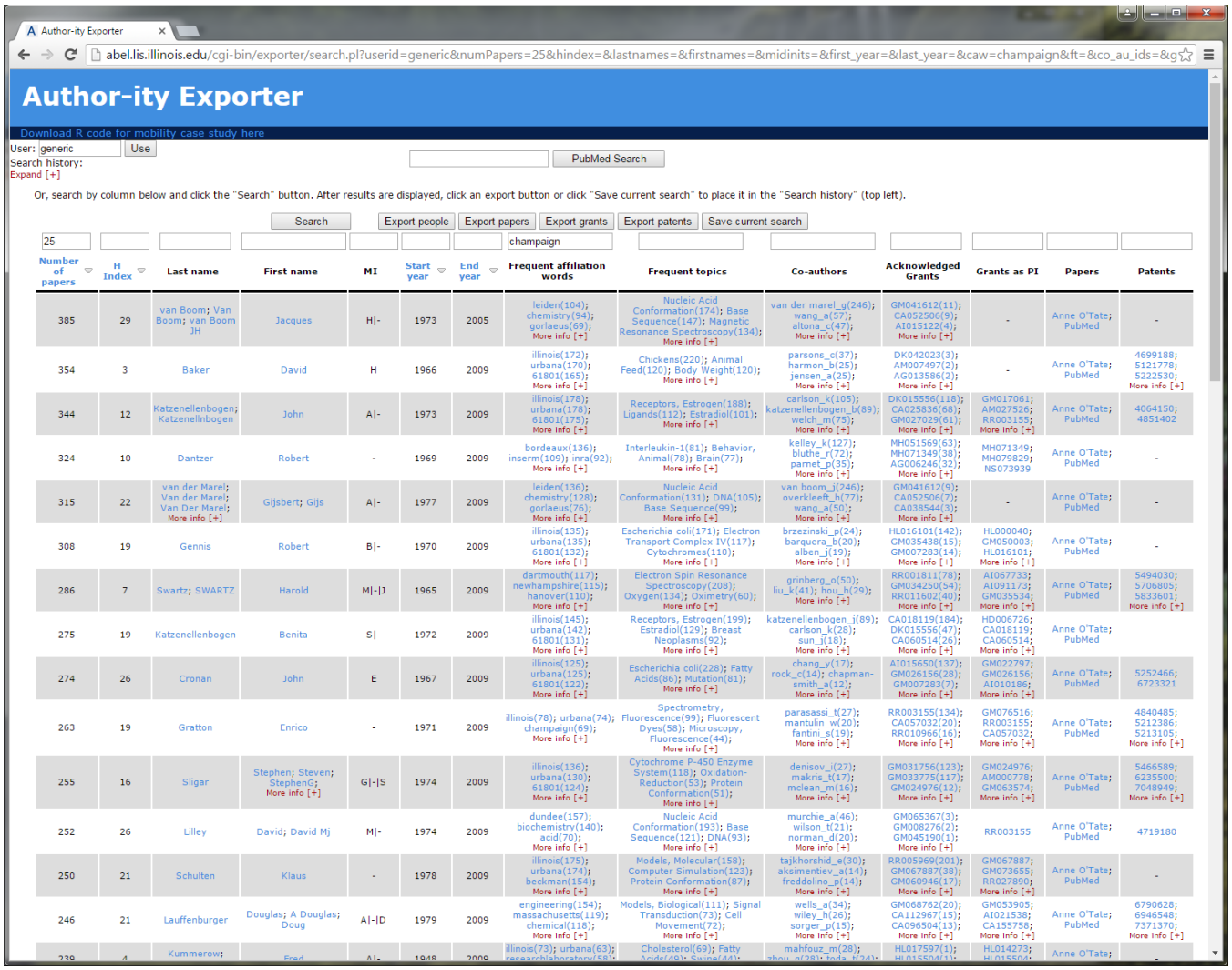

Figure 1. Screenshot of the Author-ity Exporter. Authors with at least 25 papers and "Champaign" among their top-20 most common affiliation words. The top 15 out of 785 authors, ordered by the publication count, are shown. The "Export papers" button is highlighted and permits downloading metadata on all 49,044 papers by the 785 authors.

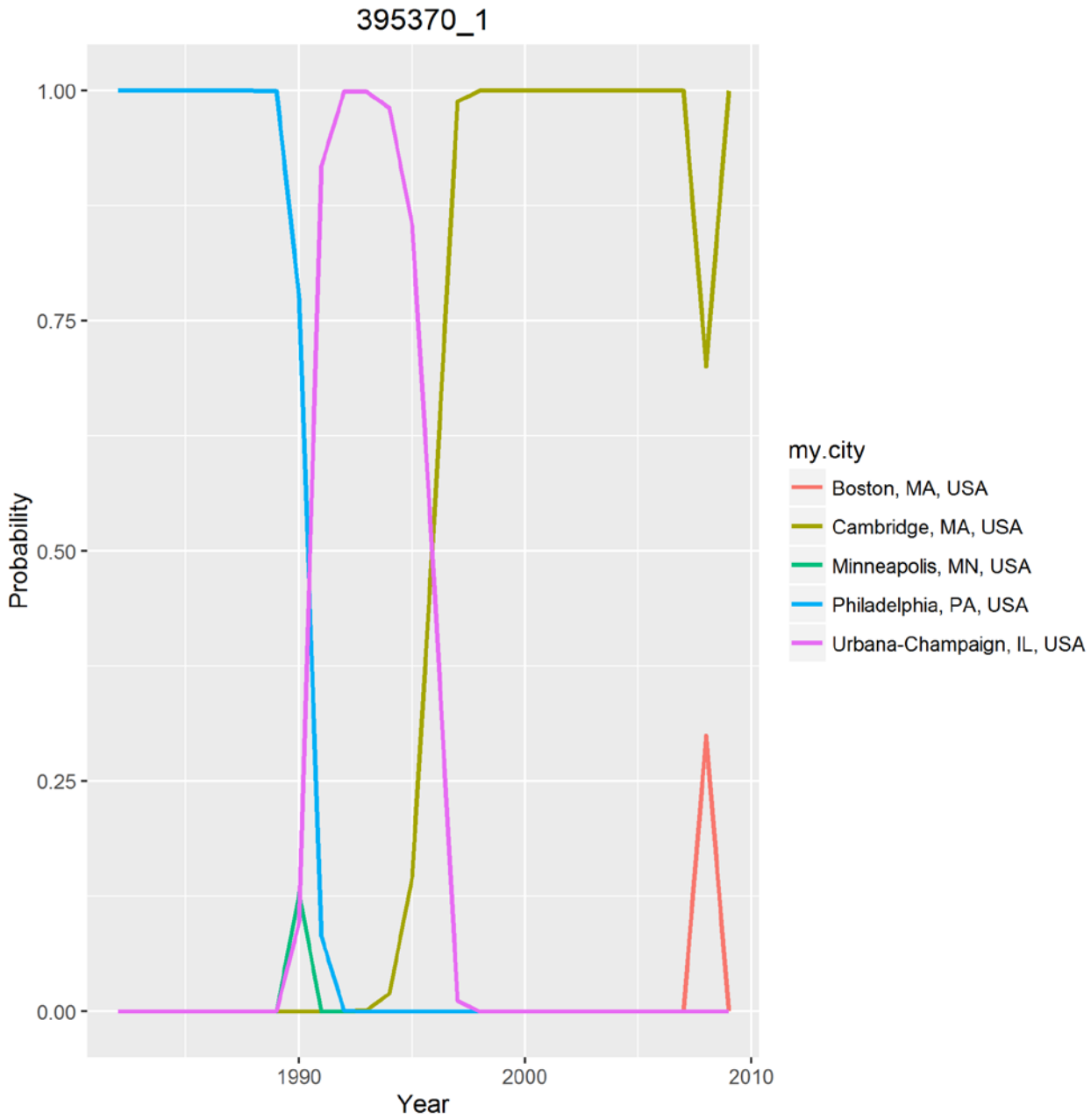

Figure 2. An illustrative example of the geo-temporal probabilistic model. The model predicts that author Douglas Lauffenburger (Author-ity ID = 395370_1) started in Philadelphia, PA in ~1980, moved to Urbana-Champaign, IL in 1991, and then to Cambridge, MA in 1996 where he remained through 2009. This aligns well with his public CV (http://web.mit.edu/dallab/people/) except that he started at UIUC in 1990 (not 1991) and MIT in 1995 (not 1996). He also completed a PhD in Minnesota in 1979 and held visiting appointments in Heidelberg, Germany in 1980 and Wisconsin in 1989-1990 both of which are not reflected in the PubMed data. Minor probability spikes occur for Minneapolis, MN in 1990 and Boston, MA in 2008. The model was able to filter them out because other cities dominated in those time periods.

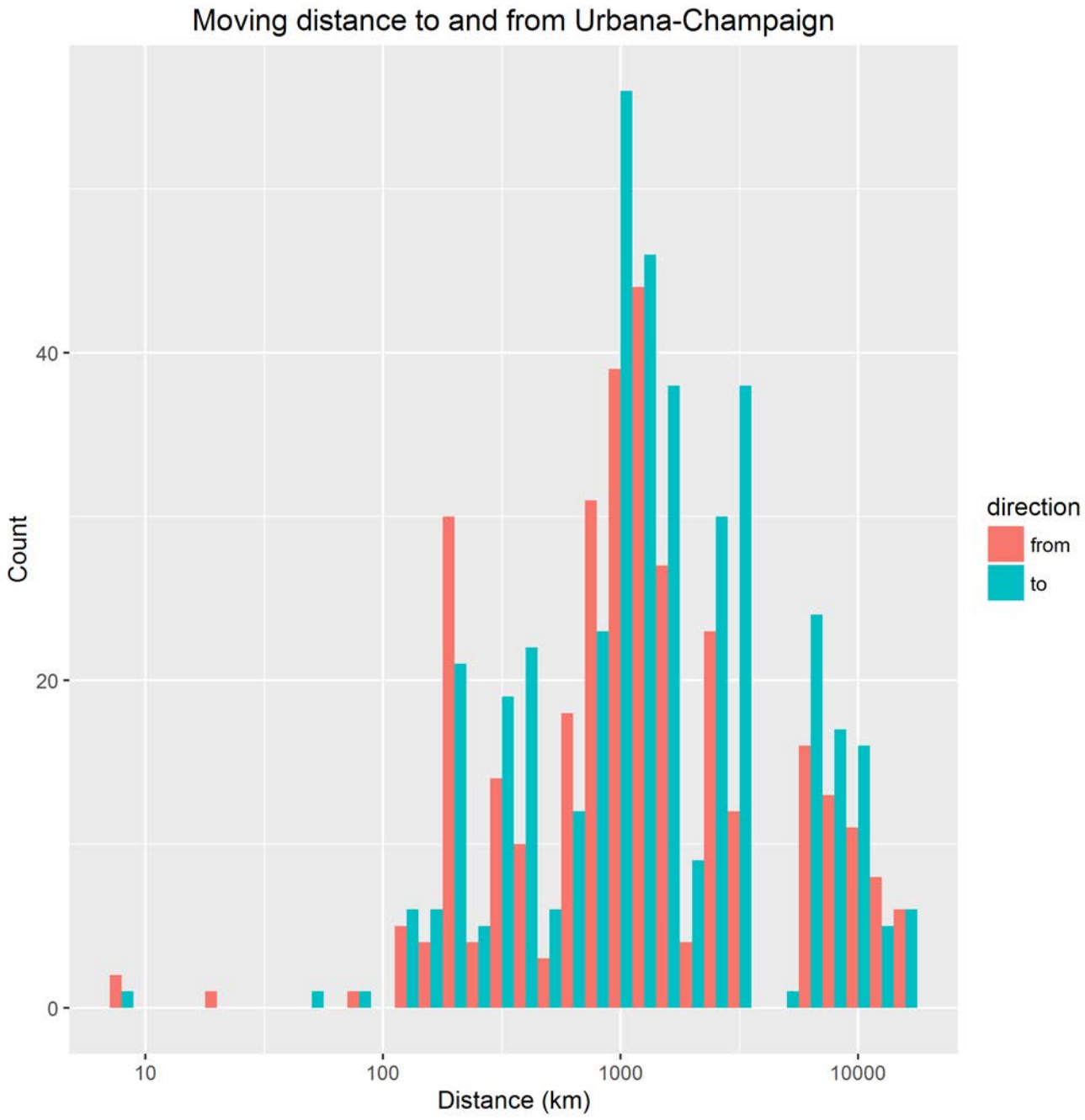

Figure 3. Distribution of distances moved by all authors into and out of Urbana-Champaign, IL. Very few move within 100km. There is a spike in moves from Urbana-Champaign to places about 200km away, and a disproportionate number of long-distance moves (> 1000km) to vs. from Urbana-Champaign.