

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 syntax¹) 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

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

Copyright © 2016. The authors.

¹ <http://www.ncbi.nlm.nih.gov/books/NBK3827/>

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.

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).

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

Here, $\text{Pr}\{city_i\}$ 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 authors², 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>

Author-ity Exporter

Download R code for mobility case study here

User: generic Use

Search history: PubMed Search

Or, search by column below and click the "Search" button. After results are displayed, click an export button or click "Save current search" to place it in the "Search history" (top left).

Search Export people Export papers Export grants Export patents Save current search

25 champaign

Number of papers	H Index	Last name	First name	MI	Start year	End year	Frequent affiliation words	Frequent topics	Co-authors	Acknowledged Grants	Grants as PI	Papers	Patents
385	29	van Boom; Van Boom; van Boom JH	Jacques	HJ	1973	2005	leiden(104); chemistry(94); goriaeus(59); More info [+]	Nucleic Acid Conformation(174); Base Sequence(147); Magnetic Resonance Spectroscopy(134); More info [+]	van der marel_g(246); wang_a(57); altona_c(47); More info [+]	GM041612(11); CA052506(9); A1015122(4); More info [+]	-	Anne O'Tate; PubMed	-
354	3	Baker	David	H	1966	2009	illinois(172); urbana(170); 61801(165); More info [+]	Chickens(220); Animal Feed(120); Body Weight(120); More info [+]	parsons_c(37); harmon_b(25); jensen_a(25); More info [+]	DK042023(3); AM007497(2); AG013586(2); More info [+]	-	Anne O'Tate; PubMed	4699188; 5121778; 5222530; More info [+]
344	12	Katzenellenbogen; Katzenellenbogen	John	AL	1973	2009	illinois(178); urbana(178); 61801(175); More info [+]	Receptors, Estrogen(188); Ligands(12); Estradiol(101); More info [+]	carlson_k(105); katzenellenbogen_b(89); welch_m(75); AG005246(32); More info [+]	DK015556(118); CA025836(68); GM027029(61); More info [+]	GM017061; AM027526; RR003155; NS073939; More info [+]	Anne O'Tate; PubMed	4064150; 4851402
324	10	Dantzer	Robert	-	1969	2009	bordeaux(136); inserm(109); inra(92); More info [+]	Interleukin-1(81); Behavior, Animal(78); Brain(77); More info [+]	kelley_k(127); bluthe_r(72); parnet_p(35); More info [+]	MH051569(63); MH071349(38); AG005246(32); More info [+]	MH071349; MH079829; NS073939; More info [+]	Anne O'Tate; PubMed	-
315	22	van der Marel; Van der Marel; Van Der Marel; More info [+]	Gijsbert; Gijs	AL	1977	2009	leiden(136); chemistry(128); goriaeus(76); More info [+]	Nucleic Acid Conformation(131); DNA(105); Base Sequence(99); More info [+]	van boom_j(246); wang_a(57); altona_c(47); wang_a(50); More info [+]	GM041612(9); CA052506(7); CA038544(3); More info [+]	-	Anne O'Tate; PubMed	-
308	19	Gennis	Robert	BI	1970	2009	illinois(135); urbana(135); 61801(132); More info [+]	Escherichia coli(171); Electron Transport Complex IV(117); Cytochromes(110); More info [+]	brzezinski_p(24); barquera_b(20); alben_j(19); More info [+]	HL016101(142); GM035438(15); GM007283(14); More info [+]	HL000040; GM050003; HL016101; More info [+]	Anne O'Tate; PubMed	-
286	7	Swartz; SWARTZ	Harold	MJ-IJ	1965	2009	dartmouth(117); newhampshire(115); hanover(110); More info [+]	Electron Spin Resonance Spectroscopy(208); Oxygen(134); Oximetry(60); More info [+]	grinberg_o(50); liu_k(41); liu_h(29); fantin_s(19); More info [+]	RR001811(78); GM034250(54); RR011602(40); More info [+]	A106733; A1091173; GM035534; More info [+]	Anne O'Tate; PubMed	5494030; 5706805; 583601; More info [+]
275	19	Katzenellenbogen	Benita	SI	1972	2009	illinois(142); urbana(142); 61801(141); More info [+]	Receptors, Estrogen(199); Estradiol(129); Breast Neoplasms(92); More info [+]	katzenellenbogen_j(89); carlson_k(28); sun_j(18); More info [+]	CA018119(184); DK015556(47); CA060514(26); More info [+]	HD006726; CA018119; CA060514; More info [+]	Anne O'Tate; PubMed	-
274	26	Cronan	John	E	1967	2009	illinois(125); urbana(125); 61801(122); More info [+]	Escherichia coli(228); Fatty Acids(86); Mutation(81); More info [+]	chang_y(17); rock_c(14); chapman-smith_a(12); More info [+]	A1015650(137); GM026156(28); GM007283(7); More info [+]	GM022797; GM026156; A1010186; More info [+]	Anne O'Tate; PubMed	5252466; 6723321
263	19	Gratton	Enrico	-	1971	2009	illinois(78); urbana(74); champaign(69); More info [+]	Spectrometry, Fluorescence(99); Fluorescent Dyes(58); Microscopy, Fluorescence(44); More info [+]	parasassi_t(27); mantulin_w(20); fantin_s(19); More info [+]	RR003155(134); CA057032(20); RR010866(16); More info [+]	GM076516; RR003155; CA057032; More info [+]	Anne O'Tate; PubMed	4840485; 5212386; 5213105; More info [+]
255	16	Sligar	Stephen; Steven; StephenG; More info [+]	GJ-IS	1974	2009	illinois(136); urbana(130); 61801(124); More info [+]	Cytochrome P-450 Enzyme System(118); Oxidation-Reduction(53); Protein Conformation(51); More info [+]	denisov_j(27); makra_t(17); mclean_m(16); More info [+]	GM031756(123); GM033775(117); GM024976(12); More info [+]	GM024976; AM000778; GM063574; More info [+]	Anne O'Tate; PubMed	5466589; 6235500; 7048949; More info [+]
252	26	Lilley	David; David Mj	MJ	1974	2009	illinois(145); dundee(157); biochemistry(140); acid(70); More info [+]	Nucleic Acid Conformation(193); Base Sequence(121); DNA(93); More info [+]	murchie_a(46); wilson_t(21); norman_d(20); More info [+]	GM065367(3); GM008276(2); GM045190(1); More info [+]	RR003155; More info [+]	Anne O'Tate; PubMed	4719180
250	21	Schulten	Klaus	-	1978	2009	illinois(175); urbana(174); beckman(154); More info [+]	Models, Molecular(158); Computer Simulation(123); Protein Conformation(87); More info [+]	tajkhorshid_e(30); aksimentiev_a(14); fredelino_p(14); More info [+]	RR005969(201); GM067887(30); GM060946(17); More info [+]	GM067887; AM000778; GM063574; More info [+]	Anne O'Tate; PubMed	-
246	21	Lauffenburger	Douglas; A Douglas; Doug	AL-ID	1979	2009	engineering(154); massachusetts(119); chemistry(118); More info [+]	Models, Biological(111); Signal Transduction(73); Cell Movement(72); More info [+]	wells_a(34); wiley_h(26); sorper_p(15); More info [+]	GM068762(20); CA112967(15); CA096504(13); More info [+]	GM053905; A1021538; CA155758; More info [+]	Anne O'Tate; PubMed	6790628; 6946548; 7371370; More info [+]
236	4	Kummerow	Erud	AL	1948	2009	illinois(73); urbana(63); missouri(58); More info [+]	Cholesterol(69); Fatty Acids(40); Squalene(41); More info [+]	mahfouz_m(28); shou_c(28); beds_f(14); More info [+]	HL017597(1); HL016581(1); More info [+]	HL014273; HL016581; More info [+]	Anne O'Tate;	-

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.

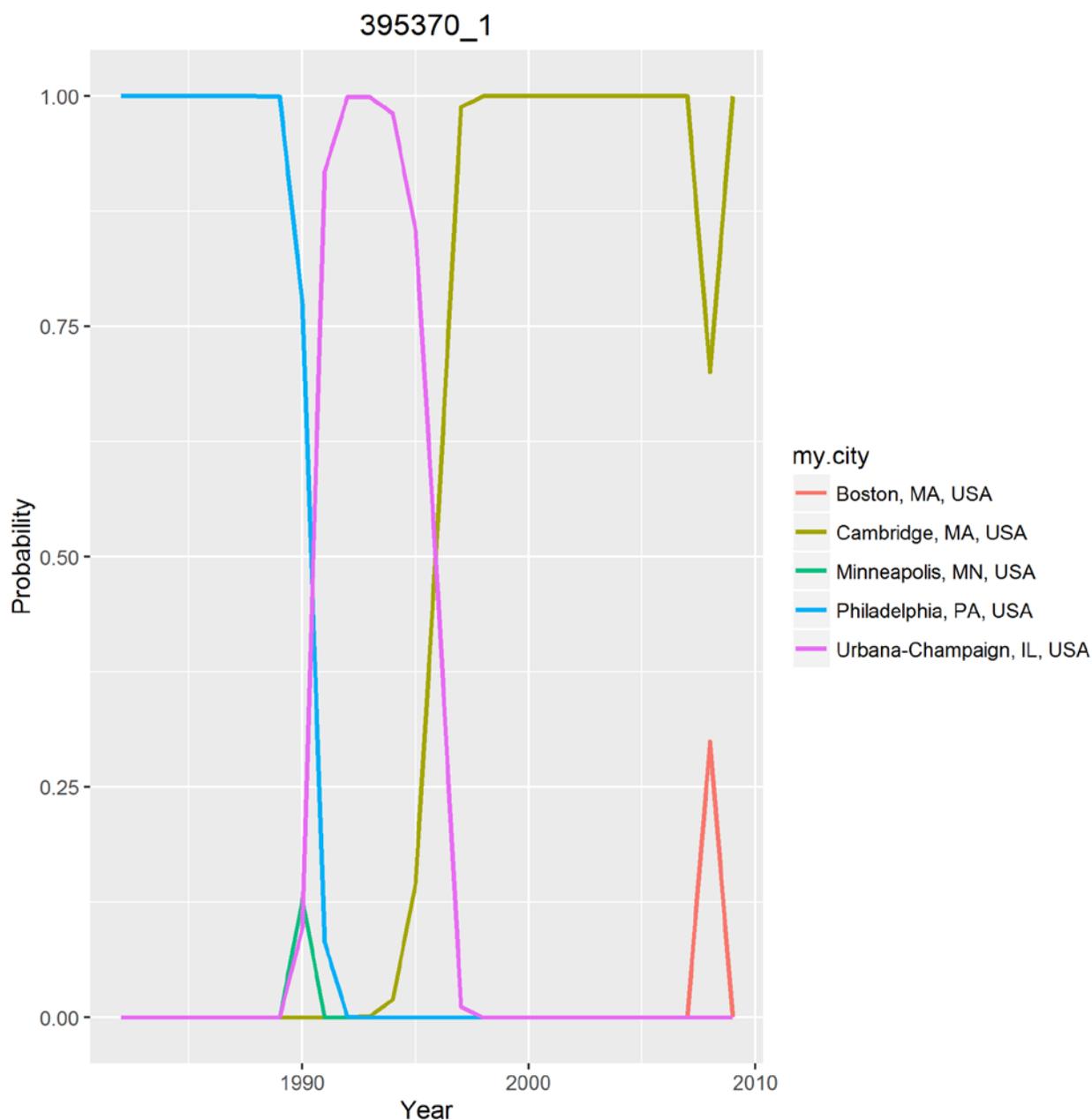


Figure 2. An illustrative example of the geo-temporal probabilistic model. The model predicts that author Douglas Lauffenburger (Author-ity ID = 395370_1) started in Philadelphia, PA in ~1980, moved to Urbana-Champaign, IL in 1991, and then to Cambridge, MA in 1996 where he remained through 2009. This 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.

Moving distance to and from Urbana-Champaign

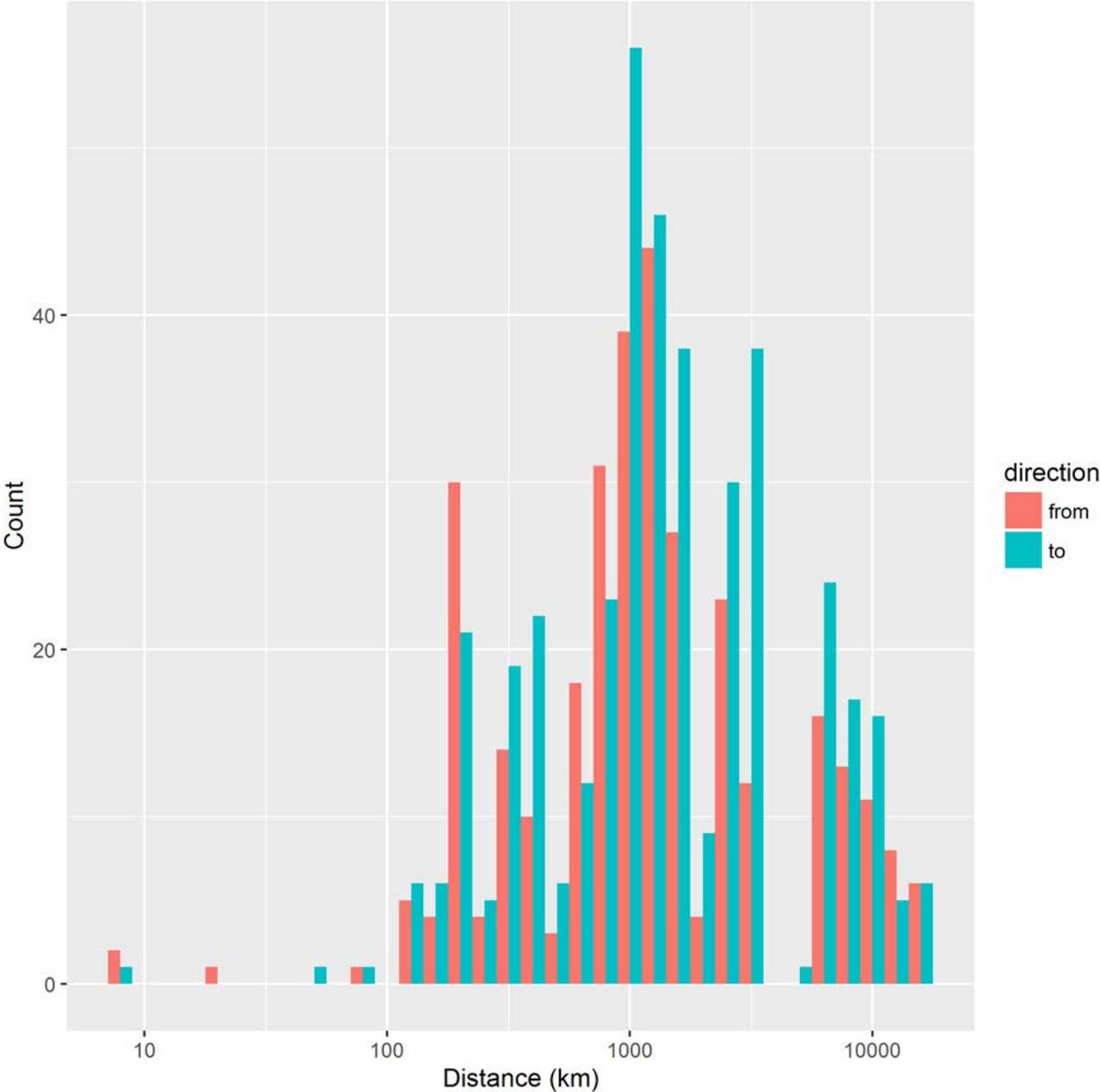


Figure 3. Distribution of distances moved by all authors into and out of Urbana-Champaign, IL. Very few move within 100km. There is a spike in moves from Urbana-Champaign to places about 200km away, and a disproportionate number of long-distance moves (> 1000km) to vs. from Urbana-Champaign.