This is a preprint of a paper accepted for publication in ACM Transactions on Knowledge Discovery from Data 2009.

Author Name Disambiguation in MEDLINE

VETLE I. TORVIK* AND NEIL R. SMALHEISER
University of Illinois at Chicago

Background: We recently described "Author-ity," a model for estimating the probability that two articles in MEDLINE, sharing the same author name, were written by the same individual. Features include shared title words, journal name, co-authors, medical subject headings, language, affiliations, and author name features (middle initial, suffix, and prevalence in MEDLINE). Here we test the hypothesis that the Author-ity model will suffice to disambiguate author names for the vast majority of articles in MEDLINE. **Methods**: Enhancements include: a) incorporating first names and their variants, email addresses, and correlations between specific last names and affiliation words; b) new methods of generating large unbiased training sets; c) new methods for estimating the prior probability; d) a weighted least squares algorithm for correcting transitivity violations; and e) a maximum likelihood based agglomerative algorithm for computing clusters of articles that represent inferred author-individuals. **Results**: Pairwise comparisons were computed for all author names on all 15.3 million articles in MEDLINE (2006 baseline), that share last name and first initial, to create Author-ity 2006, a database that has each name on each article assigned to one of 6.7 million inferred author-individual clusters. Recall is estimated at ~98.8%. Lumping (putting two different individuals into the same cluster) affects ~0.5% of clusters, whereas splitting (assigning articles written by the same individual to >1 cluster) affects ~2% of articles. **Impact**: The Author-ity model can be applied generally to other bibliographic databases. Author name disambiguation allows information retrieval and data integration to become *person-centered***,** not just *document-centered*, setting the stage for new data mining and social network tools that will facilitate the analysis of scholarly publishing and collaboration behavior. **Availability:** The Author-ity 2006 database is available for non-profit academic research, and can be freely queried via http://arrowsmith.psych.uic.edu.

Categories and Subject Descriptors: H3.3[**Information Storage and Retrieval**]: H3.3[Information Search and Retrieval]; H3.7[Digital Libraries]; I5 [**Pattern Recognition**]: I5.1[Models]; I5.2[Design Methodology]; I5.3[Clustering].
General Terms: Algorithms, experimentation, performance.
Additional Key Words and Phrases: Name disambiguation, bibliographic databases.

This research was supported by the US National Institutes of Health (NIH) grant LM008364.

Authors' addresses: Vetle I. Torvik (*corresponding author), Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel St., Champaign, IL 61820, USA; email: vtorvik@illinois.edu. Neil R. Smalheiser, Department of Psychiatry (MC912), University of Illinois at Chicago, 1601 W. Taylor St., Chicago, IL 60612, USA; email: neils@uic.edu.

1. INTRODUCTION

1.1 Motivation

Author name ambiguity in bibliographic databases has long been recognized as an important problem [e.g., Garfield 1969]. As it will be shown here for the case of MEDLINE, almost 2/3 of authors have *an ambiguous name* (where their last name and first initial is shared with one or more other authors), an ambiguous name comprises ~8 different individuals on average, and over 1/5 of authors with two or more articles have *variant name*s (where their last name, first name, or middle initial is recorded differently across their articles). Furthermore, author name searching is a major strategy employed by end users of bibliographic databases. Herskovic et al. [2007] estimated that of 2.7 million daily PubMed queries, 23% were formulated using author names exclusively. Thus, one may expect that a large portion of queries are unsuccessful because of author name ambiguity, and that author name searching would be even more common if ambiguity was not such a prevalent problem. These two points call for large scale *author name disambiguation*, a process in which one attempts to simultaneously separate cases of ambiguous names referring to different individuals and merge cases of variant names referring to the same individual, across all names and on all articles, within or across bibliographic databases [Smalheiser and Torvik 2009].

1.2 Literature review

Many different groups have tackled disambiguation as a curation or data mining problem in a variety of domains [reviewed in Smalheiser and Torvik 2009]. Fundamental approaches include: a) manual assignment by librarians [Scoville et al. 2003; MathSciNet: http://www.ams.org/mr-database/mr-authors.html]; b) community-based efforts [WikiAuthors: http://meta.wikimedia.org/wiki/WikiAuthors]; c) unsupervised clustering that groups articles by similarity [Han et al. 2005; Soler 2007; Yin et al. 2007]**;** d) supervised methods that utilize manually compiled training sets [Han et al. 2004; Reuther and Walter 2006; On et al. 2005]; and e) methods that go beyond pairwise analysis of explicit information to analyze graphs and implicit information [Bhattacharya and Getoor 2006, 2007; Huang et al. 2006; Culotta and McCallum 2006; Kalashnikov and Mehrotra 2006; Culotta et al. 2007; Galvez and Moya-Anegón 2007]. Author name

disambiguation is also closely related to several other data mining problems such as *record linkage* in administrative databases [Jaro 1995; Winkler 1995; Koudas et al. 2006], *authorship attribution* of anonymous or disputed documents using stylometry [Holmes et al. 2001; Madigan et al. 2005], and *entity resolution*, e.g., mentions of a personal name across multiple different websites [Mann and Yarowsky 2003].

We have previously shown that different articles, written by the same individual, will tend to share certain characteristic article attributes, much more so than pairs of articles authored by different individuals [Torvik et al. 2005]. We presented a probabilistic model, "Author-ity", that describes, for any two articles bearing the same name (last name, first initial), how similar the two articles are across 8 different dimensions: 1) middle initial match, 2) suffix match (e.g., Jr. or III), 3) journal name, 4) language of article match, 5) number of co-author names in common, 6) number of title words in common after preprocessing and removing *title-stopwords*, 7) number of affiliation words in common after preprocessing and removing *affiliation-stopwords*, 8) number of MeSH words in common after preprocessing and removing *mesh-stopwords*. These are calculated solely from comparing corresponding MEDLINE fields.

The resulting multi-dimensional comparison vector, or *similarity profile*, is computed for the members of two large reference or training sets – a *match set*, consisting of many (millions) pairs of articles very likely to be co-authored by the same individual across MEDLINE, and a *non-match set* consisting of many pairs of articles known to be authored by different individuals. Thus, given any pair of articles bearing the same author name (last name, first initial), we compute the similarity profile and observe its relative frequency in the match set vs. the non-match set (= the *r-value*). If the observed profile is much more frequent in the match set than in the non-match set, it is likely that the two articles were written by the same individual. The probability is easy to calculate using this method, given an estimate of the *a priori probability* of a match for that name [Torvik et al. 2005]. The pairwise model has been implemented as the Author-ity ranking tool, which is publicly accessible at http://arrowsmith.psych.uic.edu. The user inputs a specific name (last name, first initial) and is shown a list of articles bearing that name;

when the user chooses a specific article from the list, the output displays the articles ranked in descending order of probability that they were authored by the same individual.

The Author-ity model shares features of previous approaches, but is unique in several ways: It combines multiple dimensions of MEDLINE records that incorporate implicit relationships between articles as well as shared features, draws upon massive, automatically generated positive and negative training sets, incorporates nonlinear and interactive effects across dimensions, provides an explicit quantitative estimate of the probability that two articles are written by the same individual, provides a natural clustering end-point (the maximum likelihood criterion), and aims at both high precision and high recall. Although the disambiguation methods and issues are here discussed in the context of MEDLINE, they should also inform disambiguation efforts in other bibliographic databases such as CiteSeer, DBLP, arXiv, Astrophysics Data System, ACM Portal, Web of Science, Scopus, and Google Scholar.

1.3 Our hypothesis and underlying assumptions

We hypothesize that the Author-ity model, using metadata internal to the MEDLINE database, will suffice to disambiguate author names for the vast majority of articles in MEDLINE. In order to test this hypothesis, this paper extends the earlier model to describe a general approach to disambiguating author names in bibliographic databases. The enhanced model includes a) additional or corrected predictive features (first names and their variants, email addresses, and cross-correlations between specific last names and specific affiliations), b) new ways of automatically generating training sets, c) an improved method of estimating the prior probability for any given name, d) an improved algorithm for correcting transitivity violations of the form $p_{ij} + p_{jk} > 1 + p_{ik}$, and e) an agglomerative clustering algorithm that can stop at a "high precision" solution (where articles are merged with high confidence only) or continue to a natural "high recall" clustering end-point (based on the maximum likelihood criterion). Using this approach, we have generated a clustering solution for all names in the 2006 baseline version of MEDLINE.

The resulting database, named Author-ity 2006, is evaluated along several dimensions of performance, not only to assess whether the parameters of our model are optimized but also to test the following two assumptions of our model: 1) Authors' publication output is sufficiently coherent to form a single cluster containing (nearly) all of their articles yet distinguishing them from everyone else with the same name. 2) The single maximum-likelihood clustering solution adequately covers all articles in MEDLINE, old as well as new, regardless of field, geography, ethnicity, and frequency of names. In other words, we are not simply aiming at high performance, which can undoubtedly be improved by employing additional information taken from outside of MEDLINE, but rather aim to develop a model that will assist in understanding the strategies and publication behavior of scientists.

2. METHODS

2.1 The MEDLINE dataset

MEDLINE is the National Library of Medicine's (NLM) premier bibliographic database covering the most important journals in biology and medicine dating back to 1950. Each year a baseline version of MEDLINE is distributed via ftp as compressed XML files. We downloaded the 2006 baseline version, parsed the XML files and put data into two separate MySQL database tables named *Articles* and *Authors*. Each MEDLINE record is assigned a unique PubMed ID number (PMID). The *Articles* table has 15.3 million records of articles with the following fields encoded: PMID, title, journal name, author name(s), affiliation (when available), and medical subject headings (MeSH). The *Authors* table has 46.7 million records of author name occurrences, each encoded by PMID, author name position (1^{st}, 2^{nd}, 3^{rd}, etc.), last name, first initial, middle initial (when available), suffix (e.g., Jr, 2^{nd}; when available), full first name (when available), and email address (when available). An author record is uniquely identified by PMID and author name position. Author names have traditionally been encoded in MEDLINE by last name, first initial, middle initial (when available), and suffix (when available), whereas full first names have been included (when available) only since 2002. Some first names and email addresses were taken from outside MEDLINE (see Results), and the email addresses were assigned to author records using a conservative rule-based heuristic

(see Results). It should be noted that the difficulty of large scale author name disambiguation in MEDLINE arises, in part, from the fact that records often lack of primary identifying information such as affiliations, full first names and email addresses (Figure 1).

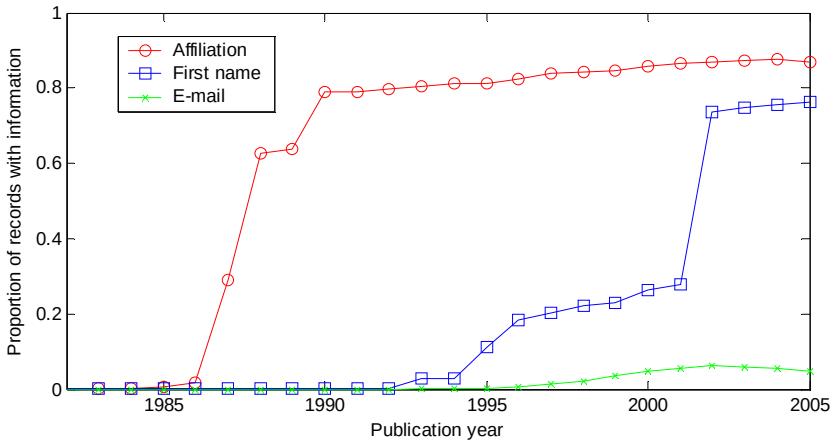

Fig. 1. Presence of affiliations, email addresses and full first names in MEDLINE records as a function of publication date. The affiliations have been included (when available) since ~1988, most often containing the affiliation of the first-listed author or corresponding author only. E-mail addresses, although rarely available, have been included in the affiliation field since ~1995. Full first names have been included (when available) since 2002. Note that some first names and emails were extracted from bibliographic sources other than MEDLINE (see Results).

Figure 2 shows the histogram of name counts across the entire 2006 baseline version of MEDLINE. The name counts follow a power law: of the 3 million unique names (based on last name and first initial), most occur only once or a few times, while a few occur very frequently. Of author names, 95% correspond to 61 articles or fewer. Only 363 (0.012% of names) contain ≥ 3,000 articles, and only 31 (0.001%) contain ≥ 8,000 articles; J. Lee has the most with 15,980 articles.

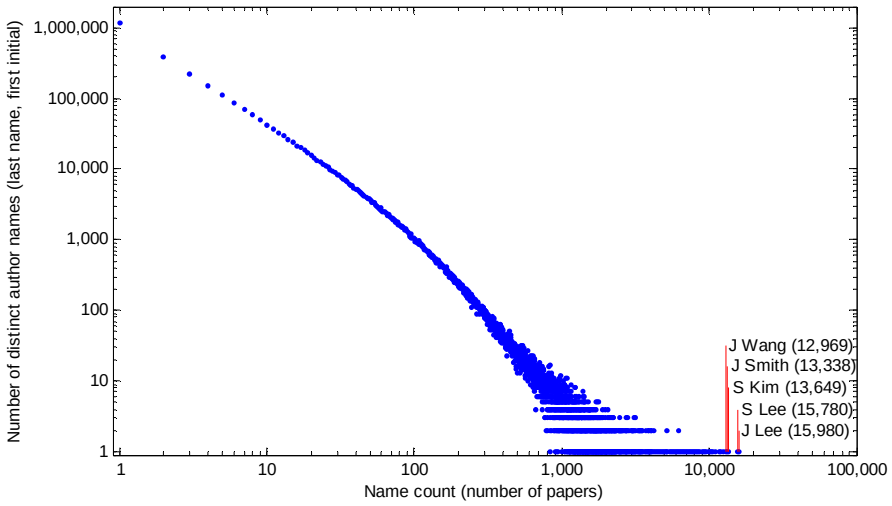

Fig. 2. Histogram of name counts in MEDLINE.

A total of 98% of articles in MEDLINE have < 10 author names listed. A relatively small number (729 articles) have over 50 authors listed, 510 of which are physics articles (appearing in *Phys. Rev. Letters* which is indexed in MEDLINE). Thus, only 219 biomedical articles have over 50 authors in MEDLINE. However, this class is worth mentioning because these are likely to be important works, particularly when clinical trials or sequencing consortia are involved, and because multi-author articles are likely to increase in the future.

2.2 Outline of the disambiguation procedure

A 5-step procedure was used to generate the disambiguated MEDLINE dataset. It will be presented here in a brief, straightforward fashion but will be discussed in more detail in the Results section.

Step 0 – "Block":

Group all the articles in MEDLINE by author name (last name, first initial) into so-called blocks. In this study, blocks include all of the articles that have a given name, but could also include articles that have highly likely name variants, share email addresses, or are

manually curated as belonging to that name. Each block is processed independently of all others by the following steps:

Step 1 – Compute similarity profiles:

For each pair of articles within the block, compute the similarity profile, a multi-dimensional vector $\mathbf{x} = \langle x_1, x_2, \ldots, x_{10} \rangle$ which is based on the different predictive features extracted from the MEDLINE records. Some features are based on attributes of the author name (first name, middle initial, and suffix) and some are related to the attributes of the article where the name occurs (title, journal, medical subject headings, co-author names, affiliation, and language). See Results for details.

Step 2 – Look up r-values:

For each pair of articles, an r-value is assigned based on the similarity profiles stored in a look-up table [Torvik et al. 2005]. For a given similarity profile **x**, r(**x**) represents the ratio of Pr{**x**|M}/Pr{**x**|N} where Pr{**x**|M} (and Pr{**x**|N}, respectively) is the probability of observing the similarity vector **x** given that the two articles are written by the same person (or different people, respectively). The r-values are pre-computed by statistical smoothing and interpolation of the training sets [Torvik et al. 2005], and stored in a look-up table. When a new similarity profile is not in the look-up table, its r-value is extrapolated from the existing similarity profiles. In two special cases, the r-value is not based on the similarity profile: **Case 1:** When the names on two different articles have the exact same e-mail address, the r-value is assigned a very high value (e.g., 10^{20}) which implies a match probability of ~1. **Case 2:** When two identical names occur on the same article, the r-value is assigned a very low value (e.g., 10^{-20}) implying a match probability of ~0. This helps to distinguish people with the same name who tend to publish together.

Step 3 – Compute a priori match probabilities and pairwise probability estimates:

The a priori probability ("prior") is defined as the proportion of pairs of articles within a given block that are written by the same person. The prior can vary dramatically from name to name (e.g., unusual names have a greater prior than common names). We use an

estimate that is tailored to a given name as follows: An initial prior is estimated using the following formula:

$$p_M = 1/(1+10^{-1.194}\, n^{0.7975}),$$

which represents the prior averaged across all names with a given number of articles *n* [see Results]. This formula is not very accurate (e.g., it severely underestimates the prior for prolific authors with an unusual name), but it provides a good starting point to convert the r-values into probabilities using the following Bayesian formula:

$$p_{ij} = 1/(1 + (1-p_M)/(p_M \times r_{ij})),$$

after which the prior is re-estimated as the proportion of pairwise probabilities > 0.5. This estimate is used together with the r-values to re-compute the pairwise probability estimates.

Step 4 – Correct pairwise probability estimates associated with triplet violations:

Given any 3 articles i, j, and k, a triplet violation is defined as a violation of the triangle inequality, at least by a small positive quantity δ (e.g., $\delta = 0.05$) [Torvik et al. 2005]:

$$p_{ij} + p_{jk} - 1 > p_{ik} + \delta \text{ and } p_{ik} \leq p_{ij}, p_{jk}$$

In this case, the lowest pairwise probability should be adjusted up, counterbalanced by adjusting the other two pairwise probabilities down [Torvik et al. 2005]. The corrected probabilities q_{ij}, q_{jk}, and q_{ik} are obtained by minimizing the weighted least squares:

$$w_{ij}(p_{ij}- q_{ij})^2 + w_{jk}(p_{jk}- q_{jk})^2 + w_{ik}(p_{ik}- q_{ik})^2.$$

Each estimate p_{ij} is weighted by the inverse of its variance, $w_{ij} = 1/(p_{ij}(1-p_{ij}))$, and the weights for the estimates below 0.5 are reduced by a factor of 0.5. This ensures that confident estimates (e.g., 0.999 and 0.001) are adjusted less in magnitude than the probabilities close to 0.5, and allows greater adjustments to the lower probabilities. The triplet correction algorithm iteratively scans through all triplets of articles to find violations. During each iteration, each pairwise probability is assigned a new probability estimate that is the average of the values indicated by all the triplet violations where it is involved. The iterations continue until there are no more violations (a maximum number of 30 iterations is allowed to ensure that the algorithm does not spend an excessive time making very small adjustments). The priors are updated once again based on the

proportion of probabilities > 0.5, and pairwise probabilities re-computed based on the original r-values and all triplet corrections repeated.

Step 5 – Carry out agglomerative clustering:

Given the set of pairwise probabilities, clustering is addressed in a maximum likelihood framework as follows:

maximize $\Pi_{i<j} [p_{ij}f_{ij} + (1-p_{ij})(1-f_{ij})]$,

subject to $f_{ij} + f_{jk} + f_{ik} \neq 2$, $\forall\ i < j < k$ and $f_{ij} \in \{0,1\}$.

Here the decision variable $f_{ij} = 1$ if the pair of articles is put in the same cluster, or = 0 if the pair is assigned to different clusters. The constraint $f_{ij} + f_{jk} + f_{ik} \neq 2$, prevents the case of labeling two out of three pairs of articles matches. This objective is approximated by a greedy agglomerative algorithm, which starts out with all articles in singleton clusters, and iteratively merges the pair of clusters with the largest average match odds. More precisely, it merges the pair of clusters c_1 and c_2 that maximizes

$\Pi_{i\in c1, j\in c2} (p_{ij}/[1-p_{ij}])/(n(c_1)n(c_2))$.

The algorithm terminates when the objective cannot be improved any more (this roughly translates to stopping when the largest pairwise probability is < 0.5).

3. RESULTS

3.1 Improvements to the original disambiguation model

The basic pairwise model [Torvik et al. 2005] was based on the 2002 baseline of MEDLINE. In this section we present additional predictive features taken from the MEDLINE records and supplemented with information extracted from publishers' pages on the internet (only the public, unrestricted face pages of online journals containing tables of contents were used for extraction).

3.1.1 Email addresses

These were extracted from the MEDLINE affiliation fields and from publishers' websites, using a heuristic algorithm to predict which author had the email address. A total of 1.46M MEDLINE records were associated with an email address, of which 270k

unique email addresses occurred on two or more MEDLINE records. These email addresses were assigned to individual author names on MEDLINE records as follows.

Condition 1 – a single author name (last name, first initial) occurs on all PMIDs associated with a unique e-mail address (this assigned 57% of the email addresses), OR

Condition 2 – if multiple names occur on all PMIDs associated with a unique e-mail address, pick the one that contains the last name within the prefix of the e-mail address (this assigned an additional 27% of the email addresses).

During this process, MEDLINE records that had multiple occurrences of the matched name (last name, first initial) were excluded. These rules capture individual emails and exclude journal or corporate emails. As a result, 228k unique email addresses were assigned to 816k author records. Personal email address match across two articles was considered absolute evidence (i.e. a gold standard) that the two articles were written by the same individual.

3.1.2 Author first names

Until 2002, MEDLINE fields did not record the first name of an author, even if that information was given in the article [NLM Technical Bulletin, Nov.-Dec. 2001]. The *Authors* table has 46.7M author name occurrences of which 8.8M (19%) contain full first names – this includes 6.1M taken from MEDLINE and 2.7M extracted from publishers' websites. In order to bring first names into the similarity profile between a pair of MEDLINE records, we first generated two new training or reference sets (in addition to the sets used for the original model): the match set consisted of ~2M pairs of names that match on email addresses. The non-match set was generated by randomly selecting 50k names and computing all pairwise comparisons with different last name and the same first initial. (Two such non-match sets were generated to check for consistency and then pooled to get estimates, resulting in 76M and 77M pairs, respectively.) Because the names in the match set are more likely to contain first name information than in the non-match set, the cases where first name was not given (i.e., $x_1 = 1$, see below) were

excluded and assigned r_1-value = 1. This left 849k pairs in the match set and 6.2M pairs in the non-match set.

Partial match on first names.

Exploring name variants in the match set, we discovered 9 general rules for partially matching first names (see x_1 = 2, 3,… , 10 below). Given two first names on a pair of articles being compared, the first name score x_1 is the greatest of the following:

11: exact match,
10: name with or without hyphen/space (jean-francois vs. jeanfrancois or jean-francois vs. jean francois),
9: hyphenated name vs. name with hyphen and initial (jean-francois vs. jean-f),
8: hyphenated name with initial vs. name (jean-f vs. jean),
7: hyphenated name vs. first name only (jean-francois vs. jean)
6: nickname match (dave vs. david)
5: one edit distance (deletion: bjoern vs. bjorn, replacement: bjoern vs. bjaern, or flip order of two characters: bjoern vs. bjeorn)
4: name matches first part of other name and length > 2 (zak vs. zakaria)
3: name matches first part of other name and length = 2 (th vs. thomas)
2: 3-letter initials match (e.g., jean francois g vs. jfg)
1: one or both names are missing,
0: otherwise.

An existing list of common nicknames was utilized [http://www.usgenweb.org/research/nicknames.shtml] and supplemented with a list of additional examples derived from examining alternative first names that were clustered together correctly during preliminary clustering (e.g., Stephan vs. Stefan).

The *Authors* table has two fields corresponding to first names - one from MEDLINE and one extracted from the publishers' websites. When there was a discrepancy between the two fields, the greater score was used. Also, in some cases, the first name field consisted of two characters that represented the first and middle initials. Therefore, when x_1 received a score of 0, and the second character of first name A matched middle initial of name B, then x_1 was assigned 1 (missing first name) instead of 0 (mismatch) and middle initial score is given a match (see x_2 = 3, below) if not already so.

The match set contained 99.0% exact first name matches ($x_1 = 11$), 0.92% partial first name matches ($x_1 = 2, 3, \ldots, 10$) and only 0.08% mismatch cases ($x_1 = 0$). In contrast, the non-match set contained 2.8% exact matches, 2.3% partial first name matches, and 94.9% mismatches. Thus, given an exact match on first name the r-value is multiplied by an average of 35 across all names (= 99/2.8) whereas when first name is a mismatch, the r-value should be multiplied by 0.0009 (= 0.08/94.9). The r-values for partial matches were estimated separately, yielding r-values ranging from 0.13 to 2.7 (see below).

Correction for first name frequency.

Very common first names, e.g., David, are more likely to match by chance than less common first names, e.g., Shawn. To estimate the effect of first name frequency, the exact matches observed in the match set and the non-match set were partitioned by first name. A random sample of first names was selected to span a wide range of counts in MEDLINE (n = 1,111 to 87,827) and the r-values were computed for each of the first names. A log-log plot of the r-values against the counts of the first names revealed a strong linear correlation (data not shown). Linear regression resulted in the following relationship: $r_1(x_1) = 10^{5.6866}\, n^{-1.0024}$ if $x_1 = 11$.

A list of first names that occur 10 or more times across all of MEDLINE was compiled, together with their counts. If a name is not on the list or occurs fewer than 10 times, the count is assigned 10 which results in an r_1-value of roughly 48,000 for very rare names. For example, David will have an r_1-value of 5.4 and Shawn = 368. In summary, the r_1-value contributed by the comparing first names is computed as follows:

$$
r_1(x_1) = \begin{cases} 10^{5.6866}\, n^{-1.0024} & \text{if } x_1 = 11 \\ 43.9 & \text{if } x_1 = 8, 9, \text{ or } 10 \\ 2.7 & \text{if } x_1 = 7, \\ 0.56 & \text{if } x_1 = 6 \\ 0.21 & \text{if } x_1 = 5 \\ 0.14 & \text{if } x_1 = 4 \\ 0.34 & \text{if } x_1 = 3 \\ 0.13 & \text{if } x_1 = 2 \\ 1 & \text{if } x_1 = 1 \\ 0.0009 & \text{if } x_1 = 0. \end{cases}
$$

3.1.3 Correction for interactive effects: name-specific correlations with affiliation words

Certain author last names were correlated closely with the author's affiliation (e.g., articles by Kim tended to come from Korea). In such cases, pairwise comparisons would frequently match on the affiliation field, and if uncorrected, this would bias the comparison scores and tend to predict too often that the two articles came from the same individual. This was corrected by testing the correlation between last name and affiliation words, for all names, and carrying out name-specific stoplisting of affiliation words (i.e., removing the contribution of those affiliation words that were correlated more than expected by chance). All last names that occurred more than 100 times in MEDLINE were examined. For each last name we collected all the affiliation words that occurred in > 30% of the records and 10 times more than expected (i.e., assuming no correlation was present). As a result, a list of 1,721 correlated (last name, affiliation word) pairs were detected such as (Wang, China), (Lee, Korea), (Lin, Taiwan), (Suzuki, Japan), and (Kumar, India). Such affiliation words are stoplisted when comparisons are made for those last names.

3.1.4 Summary of the improved pairwise model

The 10-dimensional similarity profile $\mathbf{x} = \langle x_1, x_2, \ldots, x_{10} \rangle$ is computed as follows:

x_1 = 0,1, …, or 11 first name match defined above
x_2 = 3 if middle initials match, 2 if both records' middle initials are missing, 1 if one record is missing middle initial, 0 if middle initials are different
x_3 = 1 if name suffix matches (e.g., Jr vs. Jr), 0 otherwise
x_4 = number of title words in common after preprocessing and stoplisting
x_5 = 1 if journal name matches exactly, 0 otherwise
x_6 = number of MeSH in common after stoplisting
x_7 = number matching co-authors names based on last name and both initials (includes matches with missing middle initial, e.g., JA Smith vs. J Smith counts)
x_8 = number of affiliation words in common after preprocessing and stoplisting
x_9 = 1 if both affiliations are given, 0 otherwise
x_{10} = 3 if language matches and both are non-English , 2 if both are English, 1 if one is English and the other is non-English, 0 if they don't match and both are non-English. MEDLINE records that are encoded as "undetermined language" ("und") are treated as any non-English language (i.e., are considered a match with any other non-English language).

Given the similarity profile, the r-value is computed by

$$r(\mathbf{x}) = r_1(x_1)\, r_2(x_2)\, r_3(x_3)\, r_a(x_4, x_5, x_6, x_7, x_8, x_9), r_{10}(x_{10}).$$

Where r_1 is defined above and r_2, r_3, r_a, and r_{10} were defined in [Torvik et al. 2005].

3.2 Estimating the prior probability for a given name

Given any pair of articles bearing the same author name (last name, first initial), the similarity profile is computed, and its relative frequency is observed in the match set versus the non-match set. The observed relative frequencies (= r-values) are then smoothed, interpolated, and extrapolated for profiles that were infrequently (or never) observed in the reference sets based on the flexible monotonicity criterion, which takes into account possible nonlinear and interactive effects across dimensions. If the observed profile is much more frequent in the match set than in the non-match set, it is likely that the two articles were written by the same individual [Torvik et al. 2005]. However, the r-value is insufficient for estimating the probability that a pair of articles are written by the same individual: one also needs an estimate of the a priori probability of match for the given name [Torvik et al. 2005]. For example, if the name is very unusual (e.g., D. Gajdusek), the chances are better that any two randomly chosen articles with that name are written by the same individual than if the name is very common (e.g., J. Smith).

Initially, the a priori probability was estimated by using information from articles within the training sets that share no article attributes [Torvik et al. 2005]. This proved to be inaccurate in certain cases. We therefore adopted an alternative 3-step approach: **First**, a preliminary prior probability was assigned based on the frequency of the name (last name, first initial) in MEDLINE, and **second**, after computing the pairwise probabilities for that name, the prior probability was assigned as the proportion of article pairs having pairwise probability estimates > 0.5. **Third**, using this preliminary estimate, we carried out triplet correction, after which the prior was re-estimated by the proportion of probabilities > 0.5.

The preliminary prior was estimated by first taking a random sample of ~200 names that occurred on n = 2, 3, 4 or 5 articles. These articles were manually disambiguated, and the priors averaged for each n. On average, 90% of names that only occur on two articles in

MEDLINE refer to the same person, and the proportion decreases with *n* roughly in the following fashion:

$$p_M = 1/(1+10^{-1.194}\, n^{0.7975}).$$

For example, author names with n = 100, 1,000, or 10,000 articles will have initial priors p_M of 0.28, 0.06, and 0.01, respectively. We plotted the estimated priors using the 3-step procedure described above, to confirm that this relationship roughly holds for $n > 5$ (data not shown). It should be noted that the initial prior is not of critical importance (except names with very low frequencies) because the 3-step procedure will customize the estimates to individual names.

3.3 Correcting transitivity violations

The comparison vector, or similarity profile, documents the extent to which two articles share features explicitly. However, implicit information can also be used for disambiguation. For example, an article written by J. Thompson and N. Willow does not share any co-authors with one written by J. Thompson and W. Fried. Yet, if there exists a third article by J. Thompson, N. Willow and W. Fried, it is quite likely that the same J. Thompson wrote all three articles. The triplet correction procedure quantifies the extent to which the explicit pairwise probabilities need to be adjusted in the light of another article bearing the same author name [Torvik et al. 2005]. Given any 3 articles i, j, and k, a triplet violation is defined as a violation of the triangle inequality, at least by a small positive quantity δ (e.g., 0.05):

$$p_{ij} + p_{jk} - 1 > p_{ik} + \delta, \text{ and } p_{ik} \leq p_{ij}, p_{jk}.$$

The corrected probabilities q_{ij}, q_{jk}, and q_{ik} are obtained by minimizing the weighted least squares $w_{ij}(p_{ij}- q_{ij})^2 + w_{jk}(p_{jk}- q_{jk})^2 + w_{ik}(p_{ik}- q_{ik})^2$, which has the closed form solution:

$$q_{ij} = [w_{ij}(w_{jk} + w_{ik})p_{ij} + w_{jk}w_{ik}(1 + p_{ik} - p_{jk})] / \mathit{den}$$
$$q_{jk} = [w_{jk}(w_{ij} + w_{ik})p_{jk} + w_{ij}w_{ik}(1 + p_{ik} - p_{ij})] / \mathit{den}$$
$$q_{ik} = q_{ij} + q_{jk} - 1 = [\, w_{ij}w_{jk}(p_{ij} + p_{jk} - 1) + w_{ik}\,(w_{ij} + w_{jk}\,)p_{ik}\,] / \mathit{den}$$

where *den* = $w_{ij}w_{jk} + w_{ij}w_{ik} + w_{jk}w_{ik}$.

The triplet correction algorithm iteratively scans through all triplets of articles to find violations. During each iteration, each pairwise probability is assigned a new probability that is the average of the values indicated by all the triplet violations in which it is involved. The connected components (obtained by computing the spanning tree induced by linking pairs of articles with $p_{ij} > 0.5$) are computed before each iteration, and triplet corrections are performed on each connected component separately. This dramatically improves the CPU time, especially for the later iterations, when the largest connected component is often significantly reduced. The iterations continue until there are no more violations (a maximum of 30 iterations is allowed to ensure that the algorithm does not spend an excessive time making very small adjustments.)

During initial testing it became clear that a fixed weighting scheme (i.e., $w_{ij} = w_{jk} > w_{ik} = 1$) for triplet corrections left room for improvement. For example, the probabilities for which we had high confidence (e.g., 0.999 or 0.001) were allowed to shift just as much as the less confident probabilities (e.g., 0.5). It would be better to weight an individual probability estimate by a measure of its reliability. Thus, the inverse of the variance is a natural choice, as the weighted least squares are then related to maximum likelihood estimates. Because the inverse of the variance can lead to dramatically different weights, two additional intermediate schemes were tested, the square root and log-base.

INV: Inverse of the variance: $w_{ij} = (p_{ij}(1- p_{ij}))^{-1}$
LOG: Log of INV: $w_{ij} = - \log(p_{ij}(1- p_{ij}))$
SQRT: Square root of INV: $w_{ij} = (p_{ij}(1- p_{ij}))^{-1/2}$

Each scheme was tested with options of weighting the high probabilities more heavily than the lower probabilities (w_{ij} reduced by 0.5 if $p_{ij} < 0.5$), as well as excluding corrections of very high and very low probabilities (p_{ij}, $p_{jk} > 0.95$ and $p_{ik} < 0.05$). All in all 21 different weighting schemes were tested. As gold standards, a total of 327 Community of Science (COS) [http://www.cos.com] profiles were downloaded. COS

profiles include PMIDs on their webpages, which allows for easy matching to MEDLINE records. These profiles were selected by 1) having one or more of the strings Medic, Surgery, Health, Bio, Pharm, or Chem in their expertise field, 2) the name (last name, first initial) had 4 or more profiles, and 3) the name had 100-500 articles in MEDLINE. (Very common names had to be excluded because their COS profiles often contained articles written by several different individuals. This occurred because individuals creating COS profiles were asked to select or deselect articles from a list of articles bearing the same name, and people confronted with a long list sometimes simply accepted them all.)

When two different COS profiles from different individuals having the same name (last name, first initial) were merged into a single set and then clustered, we found that the inverse variance method of triplet correction weighting (using reduced weight on lower probabilities: w_{ij} reduced by 0.5 if $p_{ij} < 0.5$) gave the best performance in separating the articles written by the two different individuals. This weighting scheme was therefore adopted generally in our project.

3.4 High-precision vs. high-recall clustering solutions

Pairwise comparisons were made for all author names on all articles in MEDLINE that share a name (last, first initial) within the author name field, and agglomerative clustering was carried out in two different ways: stopped at a high precision point ($p_{ij} = 0.95$) or at the maximum likelihood point ($p_{ij} = 0.5$). We found that high-precision clusters very often split the articles written by one individual into multiple clusters, corresponding to different groups of co-authors, different topics, different affiliations, etc. This problem was minimized with the maximum-likelihood clustering strategy, and surprisingly, we found that maximum-likelihood clusters contained relatively little "lumping" of distinct individuals into the same cluster (see below). Thus, the maximum-likelihood strategy was adopted generally in our project.

3.5 Characterization of author-individual clusters

In the Author-ity 2006 disambiguation dataset, there are 46.7M author name instances on 15.3M articles, resulting in a total of 6.7M distinct predicted author-individual clusters. About 46% of individuals have published only a single article; the average number of articles per individual is 6.9; and 95% of individuals have published 28 or fewer articles. Because a given name may include multiple individuals, the number of articles associated with a given author name is somewhat greater, but still, 95% of author names comprise 61 articles or fewer. Author-ity 2006 exhibits a high rate of author name ambiguity. Almost 2/3 (4.3M/6.7M) of individuals have an ambiguous name (defined by last name and first initial), whereas ambiguous names make up ~8 different individuals (clusters) on average. Furthermore, ~1/5 (707k/3.65M) of individuals (clusters) with two or more articles have variant first names, or variant/missing middle initials.

3.6 Evaluation of the disambiguation dataset

Evaluation of clustering across all MEDLINE authors has been a major effort - such a large, heterogeneous and evolving dataset requires a variety of different types of evaluation, covering a variety of error measures from several different perspectives. We have looked at measures of *recall* of the blocking procedure (number of pairwise comparisons in our dataset that refer to the same individual / number of pairs of articles written by the same individual across MEDLINE), *lumping* (assigning articles by different people to the same cluster), and *splitting* (assigning articles by the same person to different clusters). Each of measure can potentially be viewed from at least four different perspectives: from the perspective of individuals (or predicted author-individual clusters), author names, articles, or article pairs. For example, lumping is a type of error that will be defined and discussed in the present paper as the percentage of author-individual clusters that contain articles by 2 or more individuals (bearing in mind that lumping can also be measured as the percentage of articles that are erroneously placed into author-individual clusters, or the percentage of names that involve some degree of lumping, etc.). Similarly, we will define splitting here as the percentage of articles written by one individual that fail to be assigned to their major author-individual cluster (bearing in mind that splitting can also potentially be measured as the percentage of individuals

that have articles assigned to more than one cluster, or the percentage of names that involve some degree of splitting).

In principle, the performance of the disambiguation dataset might be limited by three different factors: The internal parameters of the model might not be fully optimized across the MEDLINE article set, or may need to be modified for particular special cases (e.g. extremely common names). Alternatively, the types of information that are incorporated into the current model so far may not suffice to give high performance in disambiguation. Finally, the underlying assumptions of the model will fail to apply to scientists whose research output is diverse. For example, Dr. Thomas H. Jobe, Professor in Psychiatry at University of Illinois at Chicago, has written articles on SPECT imaging in patients with traumatic brain injury; theoretical articles on cybernetics; and historical articles on studies of depression in the seventeenth century (and beyond MEDLINE, has published a book on the Kennedy assassination co-authored by a Mafia Don's daughter). Another real-life factor concerns closely related individuals (with the same name) who work in the same place, or in the same field. For example, Leonard A. and Leonore A. Herzenberg are husband and wife, work in the same department at Stanford, share a homepage on the internet and have published articles together. Such cases (e.g. when lacking first name information) may represent a challenge even for manual disambiguation.

3.6.1 Evaluation of performance using a random sample of articles

First, 100 names of MEDLINE authors were chosen at random, and then a pair of articles was randomly chosen for each name; these pairs were disambiguated manually, using additional information as necessary and available (e.g., author or institutional homepages, the full-text of the articles, Community of Science profiles, Google searches, etc.). Two different raters did the task separately. We found that manual disambiguation is quite difficult, and in a significant number of cases, it was not possible to be sure whether or not the two articles were written by the same individual. In a few cases, one rater said that the two articles were "definitely by different people" and the other said they were "definitely by the same person"! After resolving inter-rater differences, assignments that

were uncertain or merely "probable" were ignored, leaving 62 manual assessments which were compared against the Author-ity 2006 assignment. No articles from different individuals were placed into the same cluster, and only 1 case was found in which articles from the same person were placed in different clusters.

3.6.2 Evaluation of splitting using COS profiles, ISI highly-cited datasets, grant numbers, and self-citation datasets as gold standards

First, a set of 20,085 COS profiles was chosen at random (among those having at least 2 PMIDs; names were excluded if a PMID had multiple occurrences of the name, or if the PMID did not match the name in MEDLINE). As discussed above, it was necessary to exclude names that were associated with > 300 articles in MEDLINE. The average number of articles per COS profile was 20.0 (SD = 21.1). Bearing in mind that the evaluated COS profiles are not entirely reliable as gold standards, and may overestimate the incidence of splitting, we found that on average 98.7% (SD = 6.1%) of the articles were assigned to the largest cluster (i.e. splitting affects 1.3% of articles), and 99.9% (SD = 1.6%) of the articles were assigned to the two largest clusters. Thus, in the few cases when "splitting" did occur, it generally took the form of assigning one article to a singleton cluster. This suggests that a scientist's publication output tends to be coherent but often includes one or a few atypical articles.

Second, lists of curated publications were downloaded from the ISI Highly Cited researchers database [http://www.isihighlycited.com], including only people having > 100 articles and whose listed expertise was related to biomedicine. Each citation on a list was included according to the following rules: 1) the author name (last name, first initial) had to be in the MEDLINE record, 2) ≥ 90% of MEDLINE title words had to occur in the title of the ISI record, and they had to be in the same order. If multiple matches satisfied the rules, the ISI citation was excluded. MEDLINE records with multiple occurrences of the same name (last name, first initial) on any single article were excluded. This left 2,313 individuals with an average of 85.8 (SD = 104.7) articles. We found that on average 98.2% (SD = 7.3%) of the articles were assigned to the largest cluster (i.e.

splitting affects 1.8% of articles), and 99.5% (SD = 3.4%) of the articles were assigned to the two largest clusters.

Third, over 1.3 million MEDLINE records have grant numbers listed. We assigned standardized CRISP grant numbers to the MEDLINE records using simple mapping rules (taking into account some of the variations used by authors to report grant numbers in articles), and requiring that the name of the principal investigator (PI) from the CRISP record (retrieved from http://crisp.cit.nih.gov) be listed as an author in the MEDLINE record. Then, all grants were grouped by PI (multiple grants were identified by grant numbers co-occurring on at least one of their articles), and then we collected all articles with these grants listed. This resulted in 83,992 groups with 2 or more articles, with an average of 12.6 (SD = 21.0) articles per group. Note that a PI may correspond to multiple groups, for example, if the PI had two grants numbers that never co-occurred in an article. Within the CRISP gold standard we found that on average 99.0% (SD = 5.7%) of the articles were assigned to the largest cluster (i.e. splitting affects 1.0% of articles), and 99.9% (SD = 1.5%) of the articles were assigned to the two largest clusters.

Fourth, although MEDLINE lacks citation information, articles that are cross-listed in PubMed Central have citations listed. These citations have been mapped to the corresponding MEDLINE records and can be retrieved in batch-mode using Entrez Elink Utility (http://www.ncbi.nlm.nih.gov/entrez/query/static/elink_help.html). Citation lists were downloaded for 726k PubMed Central records. These articles collectively cited 3.53 million unique records in PubMed. Pairs of articles where one cited the other (and share an author name) was considered a potential self-citation. Some care was needed to remove cases where the name referred to different people. This set was therefore further restricted to articles that 1) had been cited only once; 2) did not have any names that appeared multiple times; and 3) were included in the 2006 baseline of MEDLINE. A total of 183,233 pairs of MEDLINE records were used to construct the gold standard by considering each shared name separately. This left 323,274 unique “self-citing” examples of the format <PMID of citing article, PMID of cited article, author name>. The self-citation gold standard was made up of pairs of articles and was therefore slightly different

from the three other gold standards that had clusters of articles grouped by individuals. In this case we measured the amount of splitting by the proportion of examples (of pairs of articles) assigned to separate clusters. Within the self-citation gold standard we found that 97.9% (316,584) examples were assigned to the same cluster (i.e., splitting affects 2.1% of the pairs of articles).

These gold standards represent four different slices of MEDLINE authors that are not necessarily typical of all authors in MEDLINE, but their results all agree insofar as the great majority (~98%) of articles belonging to a single person were assigned to the largest author-individual cluster, and most splitting consisted of assigning one article to singleton clusters.

3.6.3 Evaluation of lumping

To estimate the overall incidence of lumping, we collected all of the author-individual clusters in our dataset that contained two or more different first names not accounted for by known nicknames or name variants (this analysis was carried out before clustering of the most common names was completed, so the dataset included only names having < 3,000 articles). Of the author-individual clusters with at least one article with a full first name given, 0.8% had two or more different first names. We took a random sample of article pairs from these clusters bearing different first names, and found that about 2/3 of the cases referred to the same person. Thus, by this measure, only about 0.27% of clusters have articles by two different persons assigned to the same cluster. This does not include other cases of lumping (e.g. where the first name is not given at all or where two individuals share the same first name), but suggests that overall, lumping is a minor phenomenon in our dataset. This percentage agrees with our analysis in section 3.6.1, where a random sample of 62 pairs yielded no cases of lumping. It also agrees with a random sample of 148 COS profiles (corresponding to 65 different names, each with multiple COS profiles; and 14 articles per COS profile on average), which yielded only 1 cluster (0.7%) with articles by two different people, and in this case a single article was incorrectly assigned to a larger cluster.

3.6.4 Robustness of the clustering solution

The clustering solution produced by the model's pairwise probabilities was examined for robustness by randomly perturbing all pairwise probabilities up or down by a small amount. A sample of 12 names was randomly selected with a bias towards highly frequent names, in order to estimate robustness under a worst case scenario. The pairwise probabilities were first computed in the usual manner, but before initiating the agglomeration algorithm each probability was randomly perturbed up or down by a small quantity δ: $p\text{-}new_{ij} = p_{ij} \pm \delta$, for $i > j = 1,2, \ldots n$, where $\delta = 0.0001, 0.0005, 0.001, 0.005, 0.01$. The perturbations shuffle the order in which articles are merged; this is particularly significant when many probabilities are close in magnitude. The number of articles that were re-assigned was used as a measure of robustness. Even at the greatest perturbation (= 0.01), only 0.12% (+/-0.14) of the articles were re-assigned, and the majority of these cases consisted of a single article removed from or merging with a larger cluster. Thus, very few articles are assigned with marginal confidence, even with very common names; or stated another way, the clustering solution is very robust.

3.6.5 Confidence of assignments

In order to estimate the proportion of a person's articles that are significantly "atypical" relative to the others according to the features of our model, we examined a corpus of COS profiles that contained at least 10 MEDLINE articles (avoiding names with more than 300 articles in MEDLINE), leaving a total of 12,175 COS profiles which are regarded as gold standards. For each profile we randomly selected one article and computed the pairwise probabilities for that index article against the remaining articles listed on the same COS profile. Note that a probability estimate less than 0.5 suggests that the article is NOT authored by the same individual who wrote the index article (note, however, that our predictions are based on a clustering solution and are not precisely the same as the pairwise probability estimate). Figure 3 shows the histogram of average pairwise probability estimates. Overall, only a small proportion (1.6%) of articles had probability estimates less than 0.5, and 90% of the COS profiles contained no articles at all with $p < 0.5$. This suggests that the underlying assumption of the model – that an

individual tends to create a distinctively coherent body of work – does indeed hold in the vast majority of cases.

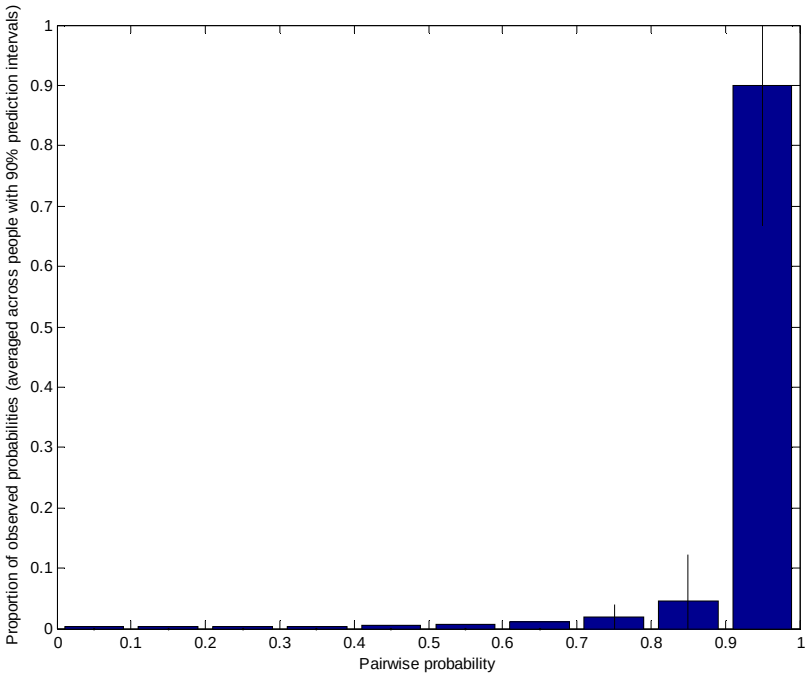

Fig. 3. Histogram of pairwise probabilities averaged over the corpus of COS profile gold standards.

3.6.6 Employing global summary parameters to characterize the performance of the model in special cases

Because manual disambiguation of the dataset is not feasible to perform on a large scale, particularly for common names, we sought an alternative "summary parameter" strategy for identifying whether the performance of the model is uniform across the dataset as a whole. The following summary parameters were examined for each author name: a) the number of predicted clusters (normalized for the number of articles associated with a given name), b) the percentage of clusters that are singletons (i.e., contain a single article), c) the size distribution of clusters (do they follow a power law?), d) the percentage of articles that are "jumpers" (i.e., that are re-assigned to different clusters

when the clustering process is carried out two different times using slightly perturbed probability estimates; see section 3.6.4), and e) the percentage of clusters that exhibit two or more first names in the same cluster (excluding known nicknames and name variants; see section 3.6.3). For example, one can ask whether highly frequent names follow the same summary parameters as do the infrequent names (which are much easier to disambiguate manually). If so, then this suggests that the dataset behaves homogeneously, and the performance exhibited by test evaluations can be extrapolated to the rest of MEDLINE.

As shown in Figures 4-6, multiple summary parameters show consistent behavior for a very wide range of name frequencies (up to 14,000 articles per name). This suggests that the probabilistic disambiguation model is, indeed, valid across MEDLINE as a whole.

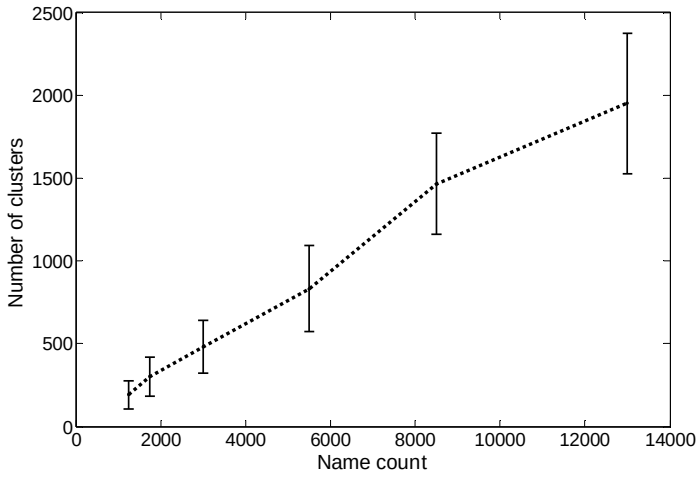

Fig. 4. Number of author-individual clusters per name as a function of the number of articles per name. Values shown are mean ± S.D.

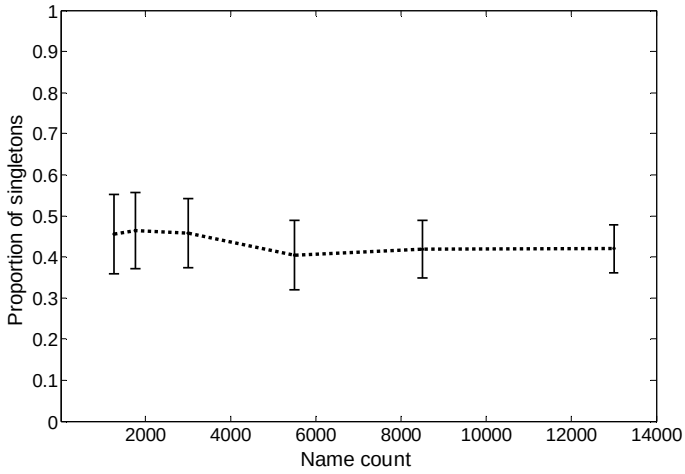

Fig. 5. Proportion of author-individual clusters containing a single article ("singletons"), as a function of the number of articles per name. Values shown are mean ± S.D.

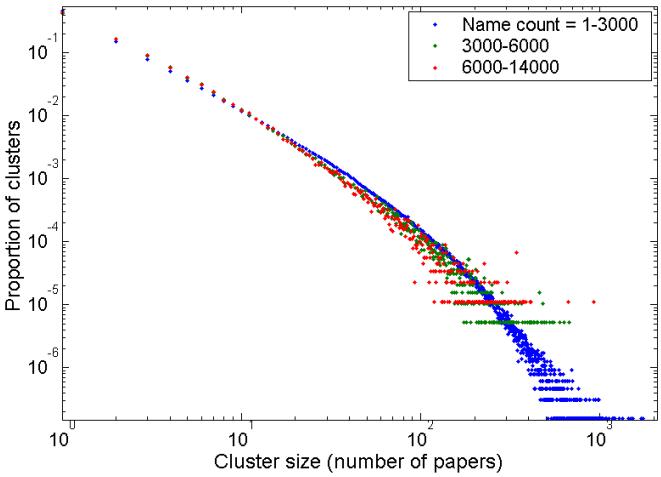

Fig. 6. Size distribution of author-individual clusters, as a function of the number of articles per cluster. The distribution is the same for rare, moderate and very common names. (There is a deviation from a true power law for authors with many articles, which probably reflects the fact that individuals have a finite lifespan so no one person can publish too many articles.)

3.6.7 Estimating the effect of blocking on recall

The current model requires that two articles must have a perfect match on author name (last name, first initial) to be disambiguated. However, occasionally the same individual's

last name is spelled differently on different articles. To estimate the effect of this blocking procedure on recall in our model, we compiled a list of all sole-authored articles in MEDLINE containing an email address that occurred on exactly two articles in MEDLINE (hence creating a gold standard of article pairs arising from the same individual). In this corpus, 1.2% of article pairs were associated with more than one last name: Of these, completely different last names were found in 10% of cases (e.g., two articles written by J. Benson vs. J. Flynn both referred to the same person, Judy Benson Flynn) whereas last name variants (1-edit distance or subsets of compound names) accounted for 90% of the differences. Among the lexical variants, 19% showed hyphen and space deletions (e.g., Olle-Goig vs. Olle Goig, or Le Roith vs. LeRoith), 24% showed 1-edit distance (deletion/addition, replacement, or character pair order flipped e.g., ae vs. ea), and 53% were cases where one name was a string subset of the other (e.g., de Lacerda vs. Lacerda or McCarter vs. Carter). Thus, the recall of this blocking procedure is estimated as 98.8% (i.e., 100% - 1.2%), and this can potentially be improved to ~99.88% if simple last name lexical variants are incorporated into the model in the future.

3.6.8 Comparing performance with other approaches

Thomson Scientific and Elsevier, two commercial competitors, maintain subscription-based bibliographic databases (Web of Science and Scopus, respectively), each containing ~ 30 million records (including MEDLINE) that have undergone author name disambiguation. Their disambiguation methods are proprietary and have not been described in detail. However, the databases represent two existing efforts to disambiguate all the records in MEDLINE, and the disambiguation datasets are available via their query interfaces. In order to compare the accuracy of the Author-ity 2006 dataset to that of Web of Science and Scopus, a random sample of 200 pairs of articles was taken from the Community of Science gold standard described above, where each pair of articles is known to be written by the same person.

A total of 61 cases had to be excluded either because the corresponding records were not found in the Web of Science or Scopus (13 cases in Web of Science and 3 cases in Scopus), or the corresponding author name had not yet been processed or lacked a high

confidence assignment in Web of Science (46 cases). Among the remaining 139 cases, the Author-ity model correctly assigned every pair to the same cluster, whereas Web of Science and Scopus split 7.8% (11/139) and 18.7% (26/139) of the pairs into separate clusters, respectively. Furthermore, in the 34 cases where Author-ity 2006 and Scopus differed, Author-ity 2006 was incorrect only once. Thus, the Author-ity model was more accurate and exhibited a lower splitting rate as assessed on this gold standard.

3.6.9 The effect of name frequency and publication date on splitting errors

We have estimated that the model split pairs of articles by the same person ~2% of the time across MEDLINE as a whole. However, author names vary widely in terms of how difficult they are to disambiguate. For example, very common names and names that occur on old articles tend to be harder to disambiguate. In order to measure the effect of publication year (x_1) and frequency of the name (x_2) on the average splitting rate (y), the self-citation gold standard dataset was used to fit the following logistic regression model.

$$y = 1/(1+\exp(-b_0 - b_1*x_1 - b_2*\log_{10}(x_2))$$

The self citation dataset was selected because it had the broadest ranges of publication dates and name frequencies of all the gold standards. The regression model fit the observed data very well. Goodness of fit was confirmed by 1) plotting the observed data against the fitted lines (not shown), and 2) refitting the model with the interaction [x_1* $\log_{10}(x_2)$] and the quadratic [x_1^2 and $\log_{10}(x_2)^2$] terms, showing that they were statistically insignificant. To check whether the curves were specific to self-citations, we separately fit a logistic regression model using the ISI gold standard. This yielded very similar curves (not shown).

Figure 7 shows the regression curves for specific years (1950 through 2000 at 10-year intervals, and 2005). The disambiguation model performs very well on the most recent articles (date of publication = 2005) where < 2% splitting occurred among names with frequency up to ~1,000, and the highest rate of splitting of 13% occurred for the most common names. In contrast, the oldest articles (publication year = 1950) are much more

difficult to disambiguate. An estimated 17% of pairs are split for moderately frequent names (1,000), whereas 2% splitting is achieved only for names with a modest frequency (~50). The high splitting rate among old articles reflects the lack of information present in the MEDLINE record.

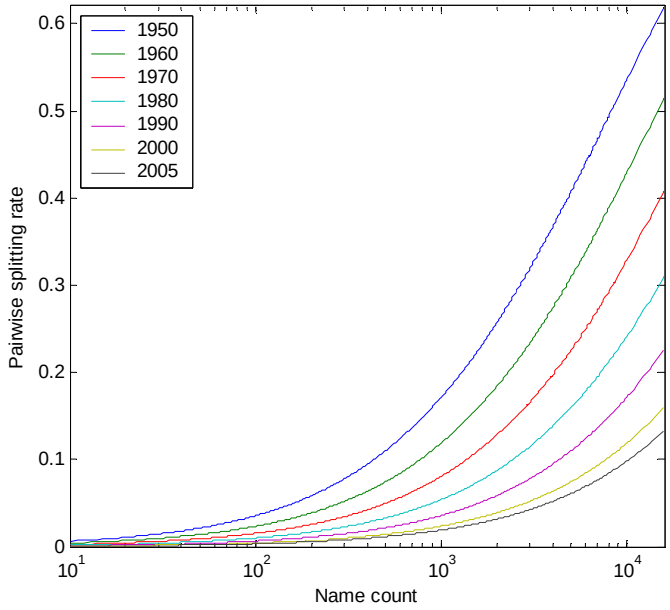

Fig. 7. Logistic regression curves showing the splitting rate (y-axis) as a function of name frequency (x-axis) and the year of publication (color coded curves) within the self-citation gold standard dataset.

To confirm our estimate of the overall prevalence of splitting across MEDLINE as a whole, we computed the estimated splitting rate for each of the 46.7 million name instances in MEDLINE based on the age of the article and name frequency using the logistic regression model. This resulted in an overall average splitting rate of 1.8% (and a 1% average for 2005). This method of estimation removes possible bias due to the having more recent articles or more names of high frequency in the self-citation dataset than MEDLINE as a whole.

It should be noted that the relative frequencies of Asian names such as Lee, Wang, and Kim have been accelerating since the mid-1980s (Figure 8). In contrast, the relative

frequency of J. Smith has decreased at a constant rate since 1970. This suggests that the disambiguation model should be further improved in the future by incorporating information regarding names written in their native languages (e.g., by publishers starting to encode Chinese pictographs alongside the English name equivalents [Qiu 2008]).

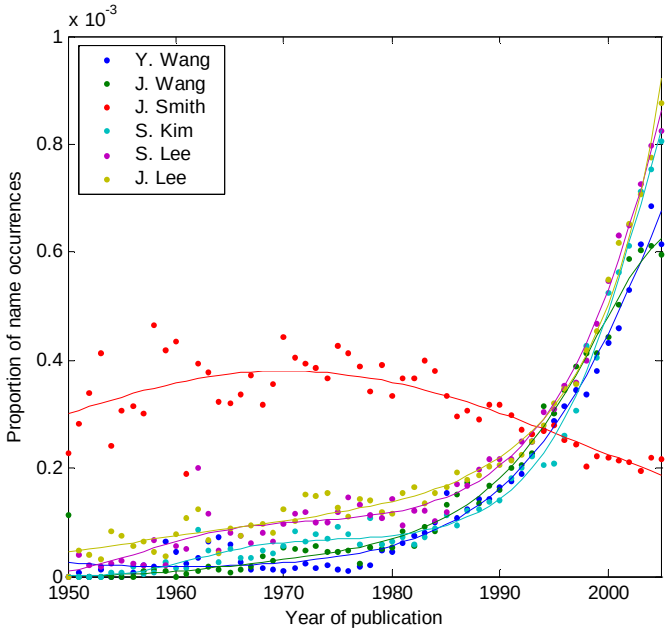

Fig. 8. Frequencies of the six most common names in MEDLINE over time. The dots correspond to the observed values, and the lines show fitted regression curves.

3.7 Scalability and computational issues

Pairwise probabilities, triplet corrections and clustering were all computed in one integrated process, requiring roughly 3 weeks to cluster all names in MEDLINE when running 8 processes (on 4 dual-CPU machines each with 4G RAM or more). MEDLINE data were extracted from XML files and stored in two MySQL tables: the *Articles* table uses 7GB of disk space for 15.3M records, and the *Authors* table uses 1.8GB of disk space for 46.7M records. All the clusters and pairwise probabilities (having values $\geq$ 0.0001; and $\geq$ 0.01 for highly frequent names) were stored in text files requiring ~250GB of disk space. A query index was generated and stored in a MySQL table, which allows

for quickly retrieving all clusters or pairwise probabilities for a given name (last name, first initial).

Initially, all programming was done in Perl, but in order to compute the high frequency names (>3,000 articles per name), memory allocation had to be controlled more closely, so the final program was implemented using a combination of Perl and C. The most frequent name (J. Lee) had 128M ($16k^2/2$) pairwise probabilities that were stored in ~500MB of RAM when using 4 bytes storage per number (typical for a floating point number in C). As a result, the computer program performed 1 million pairwise comparisons per 30 seconds. There were a total of 46.7M author name instances (on 15.3 million articles in MEDLINE), which corresponded to a total of $1.1x10^{15}$ ($= N^2/2$) possible pairwise comparisons. Restricting the pairs to ones that match on name (last name, first initial), reduced the total number of pairs to $1.2x10^{10}$, which translated into roughly 4 days CPU time. We expected that the computer program would spend most of its time on triplet corrections because each triplet iteration takes $O(N_p^{3/2})$ where N_p is the number of pairs within the block. However, in practice, the program spent roughly the same amount of CPU time on triplets as performing all pairwise comparisons for all but the most frequent names. This was due to optimizing the C code (storing as much information as possible in memory and performing as few calculations as possible within the innermost loop) and performing triplet corrections on each connected component separately (the spanning tree induced by linking pairs of articles with $p_{ij} > 0.5$).

We expect that the computational feasibility will extend to other bibliographic databases that are smaller than MEDLINE, such as CiteSeer, DBLP, arXiv, ACM Portal, and ADS, as well as databases that are twice as large (e.g., Scopus or Web of Science). The computational complexity of our approach is closely tied to the blocking procedure. The "good news" is that blocking enables parallel computations -- each block can be addressed independently of all other blocks, which in the case of MEDLINE, reduces the number of pairwise comparisons by a factor of ~100,000. The "bad news" is that the number of pairwise comparisons increases quadratically with the number of articles within each block, and the number checks for triplet violations increases cubically with

the number of articles within each block. Because our algorithms were efficient enough to process the entire MEDLINE database, we have not yet explored more restrictive blocking procedures (e.g., excluding pairwise comparisons that mismatch on middle initial for the most frequent names), or other possible speed-ups suggested for related computational problems such as adaptive blocking [Bilenko et al. 2006], sparse quadratic programming [Dominguez and Gonzalez-Lima 2006], approximate joins in databases [Koudas et al. 2006], and incremental (or online) clustering [Fisher 1987]. However, we are planning to develop algorithms to update the Author-ity database, in an online fashion, as new records are added to MEDLINE.

4. DISCUSSION

The Author-ity model [Torvik et al. 2005] was developed for estimating the probability that two articles in MEDLINE sharing the same author name are written by the same individual. The present paper extends the model to include: a) additional or corrected predictive features (first names and their variants, email addresses, and cross-correlations between specific last names and specific affiliations), b) new ways of automatically generating training sets, c) an improved method of estimating the prior probability for any given name, d) an improved algorithm for correcting transitivity violations of the form $p_{ij} + p_{jk} > 1 + p_{ik}$, and e) an agglomerative clustering algorithm that stops at a natural "high recall" clustering end-point (based on the maximum likelihood criterion). Finally, we created "Author-ity 2006", a database that assigns each name on each article in MEDLINE (Baseline 2006) to a predicted author-individual. Author-ity 2006 comprises 46.7M instances of author names appearing on 15.3M articles, and a total of 6.7M distinct predicted author-individual clusters.

About 46% of individuals in the dataset have published only a single article; the average number of articles per individual is 6.9; and 95% of individuals have published 28 or fewer articles. Almost 2/3 of individuals have an ambiguous name; conversely, an ambiguous name comprises ~8 different individuals on average.

The disambiguated dataset was evaluated along several dimensions of performance (splitting, lumping, recall, confidence, robustness, and consistency of global summary parameters) under a variety of conditions: a) using 5 different automatically generated gold standard datasets (based on email addresses, grant numbers, self-citations, ISI highly cited profiles, and Community of Science profiles); b) manually inspecting a small random sample of cases; c) comparing our approach to other approaches (by Web of Science and Scopus); d) comparing different cohorts of articles based on publication date and frequency of the name; and e) simulating random perturbation of the estimated pairwise probabilities to assess robustness.

Two fundamental assumptions of the model were tested: 1) Does an individual's publication output tend to be sufficiently coherent to form a single cluster containing (nearly) all of their articles, yet distinguishing them from everyone else with that name? 2) Does the model, which computes a single maximum-likelihood clustering solution over all articles in MEDLINE, give adequate performance across diverse cases (e.g., very old vs. very recent articles, English vs. non-English names, or very unusual vs. very common names)?

Our analyses show that the first assumption is satisfied: On average, our model captures the vast majority (~98%) of an author's articles, and for most individuals (~99.5%) distinguishes them from all other authors with the same name. The major predicted author-individual cluster corresponding to a given person fails to capture ~2% of that person's output, namely, those articles that are divergent from the others. Restricting pairwise comparisons to the same last name excludes 1.2% of articles that have variant spellings, misspellings or different last names. In other words, the *recall* of the blocking procedure is 98.8%. *Lumping* (putting two different individuals with the same name into the same cluster) affects ~ 0.5% of clusters, whereas *splitting* (assigning articles with the same name written by the same individual to > 1 cluster) affects ~ 2% of all articles. The clustering solution was also shown to be highly robust.

The second assumption also holds to a certain degree. The maximum likelihood clustering solution works well across MEDLINE as a whole, although some special cases suffer from a high rate of lumping (e.g., Japanese names that are very common yet lack middle initials), and other cases suffer from a high rate of splitting (e.g., highly frequent names or old articles).

Although the overall performance of the extended Author-ity model is excellent, there is still room for improvement. Additional predictive features could be incorporated into the model, e.g., grant numbers, self-citations, or abstracts (similarity in abstract words or "related articles" [Wilbur and Yang 1996]). As well, words used in the affiliation fields could be mapped to canonical forms of affiliations [French et al. 2000] or mapped onto a geographical system, i.e. to distinguish between words that describe countries, cities, institutions, departments, streets, etc. As shown by Tan et al. [2006] and Kanani et al. [2007], information that is external to the bibliographic record can also assist in disambiguation. For example, Tan et al. [2006] enter the title of each article as a phrase search in Google, and compile a list of the URLs that are retrieved. Two articles are compared according to how many URLs they share (weighted by various features). One could also explore modeling techniques that go beyond triplets such as co-authorship groups [Bhattacharya and Getoor 2006; Song et al. 2007], aggregate constraints [Culotta et al. 2007], or so-called collective entity resolution [Bhattacharya and Getoor, 2007].

Finally, one might extend the scope of the model (based currently on exact match of last name and first name initial) to take into account frequent spelling errors, alternative spellings of last names, compound last names, or nicknames and alternative first names that do not share first initial (e.g., Jerry vs. Gerald). Nevertheless, there will always remain a few articles that require manual disambiguation, e.g., linking articles by authors with entirely different last names due to marriage, religious conversion, or gender reassignment; linking old "stray" articles that lack information; linking articles by prolific authors who collaborate widely on diverse topics; or separating authors with the same name who work in the same place or on the same topic.

The computational complexity of our approach is closely tied to the blocking procedure. The "good news" is that blocking enables parallel computations -- each block can be addressed independently of all other blocks, which, in the case of MEDLINE, reduces the number of pairwise comparisons by a factor of ~100,000. The "bad news" is that the number of pairwise comparisons increases quadratically with the number of articles within each block, and the number checks for triplet violations increases cubically with the number of articles within each block. Because our algorithms were efficient enough to process the entire MEDLINE database, we did not explore more restrictive blocking procedures (e.g., excluding pairwise comparisons that mismatch on middle initial for the most frequent names), or other possible speed-ups suggested for related computational problems such as adaptive blocking [Bilenko et al. 2006], sparse quadratic programming [Dominguez and Gonzalez-Lima 2006], approximate joins in databases [Koudas et al. 2006], and incremental (or online) clustering [Fisher 1987]. However, we are planning to develop algorithms to update the Author-ity database, in an online fashion, as new records are added to MEDLINE.

The Author-ity 2006 database can be freely queried on the web (available at http://arrowsmith.psych.uic.edu). The user inputs an author name and is shown a list of predicted author-individual clusters ordered by the number of articles in each cluster. Each cluster has a simple summary showing the number of articles, name variants, range of publication dates, affiliation words, email address(es), topics, a link to PubMed, as well as a link to our in-house tool called Anne O'Tate [Smalheiser et al. 2008], which allows for more advanced summarization. Alternatively, the complete dataset is available upon request for non-profit academic research.

Author name disambiguation has strategic importance because it allows information retrieval and data integration to become *person-centered***,** not just *document-centered*. For example, the existing dataset should allow one to create and analyze large-scale collaboration networks in which each node represents an individual publishing in MEDLINE, and each link represents a co-authorship between two individuals. This should permit study of the factors that determine collaborations, both globally and with

reference to specific situations. Just as we have previously developed tools for identifying items and concepts that link two disparate literatures in a meaningful way [Torvik and Smalheiser 2007; Smalheiser et al. 2009], so too one can identify **individuals** who link two disparate fields of study, e.g. identifying individuals who have published in literature A and who have collaborated with individuals who have published in literature C.

In conclusion, highly scalable author name disambiguation supports the development of new data mining and social network tools, which will greatly facilitate the analysis of scholarly activity (e.g. publishing and collaboration behavior), both on the individual level and on aggregate levels (research institutes or disciplines).

ACKNOWLEDGMENTS

We sincerely thank Wei Zhang for programming assistance, Clement Yu for advice on computational issues, and Jeff Baer for permission to use Community of Science data. We also thank the US National Library of Medicine (NLM) for providing the 2006 baseline release of MEDLINE.

REFERENCES

BHATTACHARYA, I., AND GETOOR, L. 2006. A latent Dirichlet model for unsupervised entity resolution. In *Proceedings of the 6th SIAM Conference on Data Mining*, Bethesda, MD, April 2006, J. GHOSH, D. LAMBERT, D.B. SKILLICORN, AND J. SRIVASTAVA, Eds. SIAM, 47-58.

BHATTACHARYA, I., AND GETOOR, L. 2007. Collective entity resolution in relational data. *ACM Transaction on Knowledge Discovery from Data 1*, 1-36.

BILENKO, M., KAMATH, B., AND MOONEY, R.J. 2006. Adaptive blocking: learning to scale up record linkage. In *Proceedings of the IEEE Computer Society 6th International Conference on Data Mining*, Hong Kong, China, December 2006, 87-96.

CULOTTA, A., AND MCCALLUM, A. 2006. Tractable learning and inference of high-order representations. In *Proceedings of the ICML Workshop on Open Problems in Statistical Relational Learning*, Pittsburgh, PA, June 2006. Available from http://www.cs.umd.edu/projects/srl2006/proceedings.html.

CULOTTA, A., KANANI, P., HALL, R., WICK, M., AND MCCALLUM, A. 2007. Author disambiguation using error-driven machine learning with a ranking loss function. In *Proceedings of the 6th AAAI International Workshop on Information Integration on the Web*, Vancouver, CA, July 2007.

DOMINGUEZ, J., AND GONZALEZ-LIMA, MD. 2006. A primal-dual interior-point algorithm for quadratic programming. *Numerical Algorithms 42*, 1-30.

FISHER, D.H. 1987. Knowledge acquisition via incremental conceptual clustering. *Machine Learning 2*, 139–172.

FRENCH, J.C., POWELL, A., AND SCHULMAN, E. 2000. Using clustering strategies for creating authority files. *Journal of the American Society for Information Science and Technology 51*, 774-786.

GALVEZ, C., AND MOYA-ANEGÓN, F. 2007. Approximate personal name-matching through finite-state graphs. *Journal of the American Society for Information Science and Technology 58*, 1960 - 1976.

GARFIELD, E. 1969. British quest for uniqueness versus American egocentrism. *Nature 223*, 763.

HAN, H., ZHA, H., AND GILES, C.L. 2005 Name disambiguation in author citations using a K-way spectral clustering method. In *Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries*, Denver, CO, June 2005, M. MARLINO, T. SUMNER, AND F.M. SHIPMAN III, Eds. ACM, 334-343.

HAN, H., GILES, C.L., ZHA, H., LI, C., AND TSIOUTSIOULIKLIS, K. 2004. Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 4th ACM/IEEE Joint Conference on Digital Libraries* 2, Tucson, AR, June 2004, H. CHEN, H.D. WACTLAR, C.-C. CHEN, E.-P. LIM, AND M.G. CHRISTEL, Eds. ACM, 296-305.

HERSKOVIC, J.R., TANAKA, L.Y., HERSH, W., AND BERNSTAM, E.V. 2007. A day in the life of PubMed: analysis of a typical day's query log. *Journal of the American Medical Informatics Association 14*, 212-220.

HOLMES, D.I., ROBERTSON, M., AND PAEZ, R. 2001. Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities 35*, 315-331.

HUANG, J., ERTEKIN, S., AND GILES, C.L. 2006. Efficient name disambiguation for large-scale databases. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Berlin, Germany, September 2006, J. FÜRNKRANZ, T. SCHEFFER, AND M. SPILIOPOULOU, Eds. Springer-Verlag, Berlin Heidelberg, 536–544.

JARO, M.A. 1995. Probabilistic linkage of large public health data files. *Statistics in Medicine 14*, 491–498.

KALASHNIKOV, D.V., AND MEHROTRA, S. 2006. Domain-independent data cleaning via analysis of entity-relationship graph. *ACM Transactions on Database Systems 31*: 716-767.

KANANI, P., MCCALLUM, A., AND PAL, C. 2007. Improving author coreference by resource-bounded information gathering from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, January 2007, M.M. VELOSO, Ed. 429-434.

KOUDAS, N., SARAWAGI, S., AND SRIVSTAVA, D. 2006. Record linkage: similarity measures and algorithms. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, Chicago, IL, June 2006, 802-803. Supplementary tutorial slides available from *http://queens.db.toronto.edu/~koudas/docs/aj.pdf*

MADIGAN, D., GENKIN, A., LEWIS, D.D., ARGAMON, S., FRADKIN, D., AND YE L. 2005. Author identification on the large scale. Annual Meeting of the Classification Society of North America, St. Louis, MO, June 2005. Available from http://www.stat.rutgers.edu/~madigan/PAPERS/authorid-csna05.pdf

MANN, G.S., AND YAROWSKY, D. 2003. Unsupervised personal name disambiguation. In: *Proceedings of the 7th Conference on Natural Language Learning*, Edmonton, Canada, May-June 2003. Association for Computational Linguistics, Morristown, NJ, 33-40.

ON, B.W., LEE, D., KANG, J., AND MITRA, P. 2005 Comparative study of name disambiguation problem using a scalable blocking-based framework. In: *Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries* 2, Denver, CO, June 2005, M. MARLINO, T. SUMNER, F.M. SHIPMAN III, Eds. ACM, 344 – 353.

QIU, J. 2008. Scientific publishing: identity crisis. *Nature 451*, 766-767.

REUTHER, P., AND WALTER, B. 2006. Survey on test collections and techniques for personal name matching. *International Journal of Metadata, Semantics and Ontologies 1*, 89-99.

SCOVILLE, C.L., JOHNSON, E.D., AND MCCONNELL, A.L. 2003. When A. Rose is not A. Rose: the vagaries of author searching. *Medical Reference Services Quarterly 22*, 1-11.

SMALHEISER, N.R., ZHOU, W., AND TORVIK, V.I. 2008. Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. *Journal of Biomedical Discovery and Collaboration 3*, 2.

SMALHEISER, N.R., AND TORVIK, V.I. 2009. Author name disambiguation. In *Annual Review of Information Science and Technology 43*, B. CRONIN, Ed. 287-313.

SMALHEISER, N.R., TORVIK, V.I., AND ZHOU, W. 2009. Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Computer Methods and Programs Biomedicine 94*, 190-197.

SOLER, J.M. 2007. Separating the articles of authors with the same name. *Scientometrics 72*, 281-290.

SONG, Y., HUANG, J., COUNCILL, I.G., LI, J., AND GILES, C.L. 2007. Efficient topic-based unsupervised name disambiguation. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, Vancouver, BC, Canada, June 2007. E. M. RASMUSSEN, R. R. LARSON, E. TOMS, AND S. SUGIMOTO, Eds. ACM, 342-351.

TAN, Y.F., KAN, M.Y., AND LEE, D. 2006. Search engine driven author disambiguation. In: *Proceedings of the 6th ACM/IEEE Joint Conference on Digital Libraries*, Chapel Hill, NC, June 2006, G. MARCHIONINI, M.L. NELSON, AND C.C. MARSHALL, Eds. ACM, 314 - 315.

TORVIK, V.I., WEEBER, M., SWANSON, D.R., AND SMALHEISER, N.R. 2005. A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology 56*, 140-158.

TORVIK, V.I., AND SMALHEISER, N.R. 2007. A quantitative model for linking two disparate sets of articles in MEDLINE. *Bioinformatics* 23, 1658-1665.

WILBUR, W.J., AND YANG, Y. 1996. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Computers in Biology and Medicine 26*, 209-222.

WINKLER, W.E. 1995. Matching and record linkage. In *Business Survey Methods*, B.G. COX ET AL., Eds. Wiley, New York, 355-384.

YIN, X., HAN, J., AND YU, P.S. 2007. Object distinction: Distinguishing objects with identical names by link analysis. In *Proceedings of the IEEE 23rd International Conference on Data Engineering*, Istanbul, Turkey, April 2007. IEEE, 1242-1246.

Received July 2007; revised February 2009; accepted March 2009.