DE GRUYTER
OPEN

Research Paper

A Multi-match Approach to the Author Uncertainty Problem

Stephen F. Carley[1†], Alan L. Porter[1,2], Jan L. Youtie[3]

[1]Search Technology, Inc, Norcross, GA 30092, USA
[2]Georgia Institute of Technology, Atlanta, GA 30308, USA
[3]Enterprise Innovation Institute, Georgia Institute of Technology, Atlanta, Georgia 30308, USA

Citation: Stephen F. Carley, Alan L. Porter, Jan L. Youtie (2019). A multi-match approach to the author uncertainty problem. Journal of Data and Information Science, 4(2), 1–18

DOI: 10.2478/jdis-2019-0006

Received: Dec. 5, 2018
Revised: Dec. 20, 2018
Accepted: Jan. 8, 2019

Abstract

Purpose: The ability to identify the scholarship of individual authors is essential for performance evaluation. A number of factors hinder this endeavor. Common and similarly spelled surnames make it difficult to isolate the scholarship of individual authors indexed on large databases. Variations in name spelling of individual scholars further complicates matters. Common family names in scientific powerhouses like China make it problematic to distinguish between authors possessing ubiquitous and/or anglicized surnames (as well as the same or similar first names). The assignment of unique author identifiers provides a major step toward resolving these difficulties. We maintain, however, that in and of themselves, author identifiers are not sufficient to fully address the author uncertainty problem. In this study we build on the author identifier approach by considering commonalities in fielded data between authors containing the same surname and first initial of their first name. We illustrate our approach using three case studies.

Design/methodology/approach: The approach we advance in this study is based on commonalities among fielded data in search results. We cast a broad initial net—i.e., a Web of Science (WOS) search for a given author's last name, followed by a comma, followed by the first initial of his or her first name (e.g., a search for 'John Doe' would assume the form: 'Doe, J'). Results for this search typically contain all of the scholarship legitimately belonging to this author in the given database (i.e., all of his or her true positives), along with a large amount of noise, or scholarship not belonging to this author (i.e., a large number of false positives). From this corpus we proceed to iteratively weed out false positives and retain true positives. Author identifiers provide a good starting point—e.g., if 'Doe, J' and 'Doe, John' share the same author identifier, this would be sufficient for us to conclude these are one and the same individual. We find email addresses similarly adequate—e.g., if two author names which share the same surname and same first initial have an email address in common, we conclude these authors are the same person. Author identifier and email address data is not always available, however. When this occurs, other fields are used to address the author uncertainty problem.

† Corresponding author: Stephen F. Carley (E-mail: stephen.carley@searchtech.com).

Research Paper

Commonalities among author data other than unique identifiers and email addresses is less conclusive for name consolidation purposes. For example, if 'Doe, John' and 'Doe, J' have an affiliation in common, do we conclude that these names belong the same person? They may or may not; affiliations have employed two or more faculty members sharing the same last and first initial. Similarly, it's conceivable that two individuals with the same last name and first initial publish in the same journal, publish with the same co-authors, and/or cite the same references. Should we then ignore commonalities among these fields and conclude they're too imprecise for name consolidation purposes? It is our position that such commonalities are indeed valuable for addressing the author uncertainty problem, but more so when used in combination.

Our approach makes use of automation as well as manual inspection, relying initially on author identifiers, then commonalities among fielded data other than author identifiers, and finally manual verification. To achieve name consolidation independent of author identifier matches, we have developed a procedure that is used with bibliometric software called VantagePoint (see www.thevantagepoint.com). While the application of our technique does not exclusively depend on VantagePoint, it is the software we find most efficient in this study. The script we developed to implement this procedure is designed to implement our name disambiguation procedure in a way that significantly reduces manual effort on the user's part. Those who seek to replicate our procedure independent of VantagePoint can do so by manually following the method we outline, but we note that the manual application of our procedure takes a significant amount of time and effort, especially when working with larger datasets.

Our script begins by prompting the user for a surname and a first initial (for any author of interest). It then prompts the user to select a WOS field on which to consolidate author names. After this the user is prompted to point to the name of the authors field, and finally asked to identify a specific author name (referred to by the script as the primary author) within this field whom the user knows to be a true positive (a suggested approach is to point to an author name associated with one of the records that has the author's ORCID iD or email address attached to it).

The script proceeds to identify and combine all author names sharing the primary author's surname and first initial of his or her first name who share commonalities in the WOS field on which the user was prompted to consolidate author names. This typically results in significant reduction in the initial dataset size. After the procedure completes the user is usually left with a much smaller (and more manageable) dataset to manually inspect (and/or apply additional name disambiguation techniques to).

Research limitations: Match field coverage can be an issue. When field coverage is paltry dataset reduction is not as significant, which results in more manual inspection on the user's part. Our procedure doesn't lend itself to scholars who have had a legal family name change (after marriage, for example). Moreover, the technique we advance is (sometimes, but not always) likely to have a difficult time dealing with scholars who have changed careers or fields dramatically, as well as scholars whose work is highly interdisciplinary.

Practical implications: The procedure we advance has the ability to save a significant amount of time and effort for individuals engaged in name disambiguation research, especially when the name under consideration is a more common family name. It is more effective when match field coverage is high and a number of match fields exist.

Originality/value: Once again, the procedure we advance has the ability to save a significant amount of time and effort for individuals engaged in name disambiguation research. It combines preexisting with more recent approaches, harnessing the benefits of both.

Findings: Our study applies the name disambiguation procedure we advance to three case studies. Ideal match fields are not the same for each of our case studies. We find that match field effectiveness is in large part a function of field coverage. Comparing original dataset size, the timeframe analyzed for each case study is not the same, nor are the subject areas in which they publish. Our procedure is more effective when applied to our third case study, both in terms of list reduction and 100% retention of true positives. We attribute this to excellent match field coverage, and especially in more specific match fields, as well as having a more modest/manageable number of publications.

While machine learning is considered authoritative by many, we do not see it as practical or replicable. The procedure advanced herein is both practical, replicable and relatively user friendly. It might be categorized into a space between ORCID and machine learning. Machine learning approaches typically look for commonalities among citation data, which is not always available, structured or easy to work with. The procedure we advance is intended to be applied across numerous fields in a dataset of interest (e.g. emails, coauthors, affiliations, etc.), resulting in multiple rounds of reduction. Results indicate that effective match fields include author identifiers, emails, source titles, co-authors and ISSNs. While the script we present is not likely to result in a dataset consisting solely of true positives (at least for more common surnames), it does significantly reduce manual effort on the user's part. Dataset reduction (after our procedure is applied) is in large part a function of (a) field availability and (b) field coverage.

Keywords Name disambiguation; Author identifiers; Multi-match approach

1 Introduction

The ability to isolate scholarship belonging to individual authors is fundamental to assessing productivity, mobility, collaboration and scientific impact. For metrics to have meaning it is essential to build a body of scholarship unique to the subject(s) of one's study. The problem we face with an inability to successfully disambiguate author names is inaccurate and inactionable results—results that might be described as noisy at best and fraudulent at worst. Shin et al. (2014) note that addressing the author uncertainty problem is important "especially in digital libraries that are becoming more person-centric than document-centric." Several undesirable outcomes result from the failure to successfully distinguish between individual authors. Among these, Han and colleagues (2004) note this "can affect the quality of scientific data gathering, can decrease the performance of information retrieval and web search, and can cause the incorrect identification of and credit attribution to authors." Co-authorship networks are influenced as well. Diesner and Kim (2016) show that a failure to correctly disambiguate names will "misrepresent statistical

properties of co-authorship networks: It deflates the number of unique authors, number of components, average shortest paths, clustering coefficient, and assortativity, while it inflates average productivity, density, average co-author number per author, and largest component size." Negative results are not limited to the author level; they can also distort department and institution-level analyses as well.

The author uncertainty problem has become a more pressing challenge over time. The sizeable increase in scholarship from countries whose citizens share common surnames, like China (Zhou & Leydesdorff, 2006) and Korea, significantly contributes to this difficulty. The proliferation of common names (and especially anglicized Asian names) among multiple authors, variations in name spelling (for a given author's name) and prodigious scholarship (by individual authors) make name disambiguation all the more problematic. Standardized signatures shared by a myriad of authors, i.e. the namesake problem, such as a surname followed by a first initial, further complicate the plot. MacRoberts and MacRoberts (1989), as well as Smalheiser and Torvik (2009), provide useful overviews of the reasons why name disambiguation is as challenging as it is.

A number of solutions have been proposed for addressing the author uncertainty problem. Han and colleagues (2004) advance "two supervised teaming approaches to disambiguate authors in the citations." Song et al. (2007) promote a topic-based approach. Cota and colleagues (2007) advocate a hierarchical clustering technique. More recently, Shin et al. (2014) propose a graph based approach to the name disambiguation problem, whereby "where a graph model is constructed using the co-author relations, and author ambiguity is resolved by graph operations such as vertex (or node) splitting and merging based on the co-authorship." Hussein and Asgar (2017) provide a convenient survey of more recent author name disambiguation techniques. Among proposed solutions to date, machine learning, a field of computer science that focuses on teaching computer systems to learn, has emerged prominently. In this context, learning is synonymous with the ability to increasingly improve the performance of a given task (e.g., author name disambiguation). Lang and colleagues (2016) note that this technique can be dichotomized into dual method types: supervised and unsupervised. While the former predicts outcomes based on input characteristics and data, the latter cluster data based on identifying patterns from the data's characteristics (which the researcher can classify after running the algorithm).

Ferreira and colleagues (2012) conducted a survey of automatic methods for author name disambiguation to find that "the majority of the surveyed methods perform disambiguation by comparing citation records using some type of similarity function." In their analysis, a "major gap in the field is the lack of direct comparisons

among the methods under the same circumstances: e.g., same collections (e.g., many methods used different versions of collections such as DBLP)."[1] Another drawback of this approach is that when training data scarcity makes pattern detection difficult. Automatic methods will be revisited later. In the interim, however, we turn our attention to author identifier approaches to name disambiguation, which provide a key starting point for the method advanced in this study.

2 Unique Author Identifiers

Prominent among proposed solutions to author uncertainty are author identifiers. In October of 2012, ORCID (Open Researcher and Contributor ID) was launched as an open-access database to identify individual scholars. It is self-described as "a persistent digital identifier that distinguishes you from every other researcher and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between you and your professional activities ensuring that your work is recognized."[2] Authors are assigned an ORCID iD consisting of 16 characters. While some organizations require the adoption of an ORCID identifier, others do not, making coverage a function of organizational policy or the part of the world to which a given scholar belongs (Youtie et al., 2017).

Since 2008 the Web of Science (WOS), provided by Clarivate Analytics, has issued its own unique author identifier, which it calls ResearcherID. The website[3] for this identifier prompts authors to register with ResearcherID and link their publications with their database. If a given author has a preexisting ORCID iD he or she can link that to his or her ResearcherID (and vice versa). As of October 2017, more than 270,000 researchers have signed up for a ResearcherID[4] and 9,073,149 records indexed on WOS have a ResearcherID attached.[5] Given the fact that registration for this identifier is optional, it should not come as a surprise that ResearcherID coverage is less than 100%.

The database Scopus, provided by Elsevier, issues unique author identifiers which it refers to as 'Scopus IDs' to every author indexed in its dataset. Scopus was launched in 2004 and claims to draw from more than 5,000 publishers to index

[1] DBLP, which originally stood for DataBase systems and Logic Programming, is a scholarly digital library which was launched at the University of Trier, Germany, in 1993. It tracks all major computer science journals.
[2] See https://orcid.org
[3] See http://www.researcherid.com
[4] See https://clarivate.com/products/researcherid
[5] As all ResearcherIDs begin with a letter this figure is obtained by the WOS author identifier search: "A* OR B* OR C* OR D* OR E* OR F* OR G* OR H* OR I* OR J* OR K* OR L* OR M* OR N* OR O* OR P* OR Q* OR R* OR S* OR T* OR U* OR V* OR W* OR X* OR Y* OR Z*"

69 million items and 12 million author profiles.[6] Unlike ORCID iDs and ResearcherIDs, authors do not have to create their own profile to obtain a Scopus ID. It's noted, however, that occasionally a given scholar is unintentionally assigned more than one Scopus ID (as is the case with one of the authors on this paper). When this happens, the author has the option to request that each of his or her Scopus IDs be merged into one. In the absence of such a request, however, searching for an author with multiple Scopus IDs using just one of their identifiers will produce incomplete results. Scopus author profiles can be linked to the same author's ORCID account (and vice versa). Gasparyan and colleagues (2017) provide an overview of author identifier approaches.

While author identifiers are a truly valuable response to the author uncertainty problem, in and of themselves they do not provide a comprehensive solution for several reasons. Given that not all authors have applied for, or been issued, author identifiers, a number of authors do not possess unique identifiers. Among those that do, coverage is oftentimes less than robust. An author on this paper (Alan Porter) has a publication count of 234 on WOS (as of late 2017), but a search for his work using his ORCID iD or ResearcherID results in 40% of scholarship belonging to him.[7] Youtie and colleagues (2017) find that 19% of all WOS documents published between 2000 and 2016 are associated with one or more ORICD iDs. Author identifiers are, without doubt, a huge step in the right direction. In and of themselves, however, they do not provide a comprehensive solution to the author uncertainty problem. In this study we seek a more holistic approach.

3 The Use of Author Identifiers in Conjunction with Other Fielded Data

How best then to address author uncertainty? The method we propose uses author identifiers as a first step in the name consolidation process. Our approach is deductive in nature—we begin with a large dataset consisting of a given author's true positives along with a significant amount of noise. The starting point is a WOS search for a given author's surname, followed by a comma, followed by the first initial of his or her first name—e.g., a WOS search for author John Doe would assume the form: 'Doe, J'. This search results in all authors with last name 'Doe' and first initial 'J'. Some of these authors—e.g., Doe, Jane—are clearly not the John Doe we seek. The assignment of other author names is less straightforward, however. For example, in our search results we see the following: (i) Doe, J, (ii) Doe, JE, (iii) Doe, J E, (iv)

[6] See https://www.elsevier.com/solutions/scopus/content

[7] The same author has been issued no fewer than eight Scopus IDs, making a search for his scholarship on SCOPUS problematic as well.

Doe, John and (v) Doe, John E. It is not immediately apparent which of these refers to the John Doe we seek. By way of comparison, one of the authors on this paper has more than 200 records indexed on WOS, and, among these, has the following variations in name spelling: (i) PORTER, AL, (ii) Porter, Alan L, (iii) Porter, Alan, (iv) Porter, A and (v) Porter, A L. While we know with certainty that each of the proceeding five names refers to our colleague and co-author, the same conclusion is more difficult to arrive at in the case of a John Doe who we've never met or interacted with.

Author identifiers offer assistance for the conundrum we face in some instances, but not others. In the case of John Doe, for instance, 13% of the search results for Author = 'Doe, J' are assigned an ORCID iD. More to the point, just one of the five preceding names that might possibly refer to the John Doe we're interested in is assigned an ORCID iD in WOS search results (Doe, John). It is not currently possible to know, on the basis of ORCID identifiers alone, whether 'Doe, John' and 'Doe, J' refer to one and the same person. Given coverage issues, we must rely on additional approaches for isolating the scholarship of the John Doe we seek.

The approach advanced in this study is based on commonalities among author data in search results. We cast a broad initial net—i.e., a WOS search for a given author's last name, followed by a comma, followed by the first initial of his or her first name (e.g., 'Doe, J'). Results for this search typically contain all of the scholarship legitimately belonging to this author in the given database (i.e., all of his or her true positives), along with a large amount of noise, or scholarship not belonging to this author (i.e., a large number of false positives). From this corpus we proceed to iteratively weed out false positives and retain true positives. Author identifiers provide a good starting point—e.g., if 'Doe, J' and 'Doe, John' share the same author identifier, that is sufficient for us to conclude these are one and the same individual. We find email addresses similarly adequate—e.g., if two author names which share the same surname and same first initial have an email address in common, we conclude these authors are the same person. As previously noted, however, author identifier datum is not always available. The same holds true for email addresses as well. When this occurs, other fields are used to address the author uncertainty problem.

Commonalities among author data other than unique identifiers and email addresses is less conclusive for name consolidation purposes. For example, if 'Doe, John' and 'Doe, J' have an affiliation in common, do we conclude that these names belong the same person? They may or may not; affiliations have employed two or more faculty members sharing the same last and first initial. Similarly, it's conceivable that two individuals with the same last and first initial publish in the same journal, publish with the same co-authors, and/or cite the same references. Should we then

ignore commonalities among these fields and conclude they're too imprecise for name consolidation purposes? It is our position that such commonalities are indeed valuable for addressing the author uncertainty problem, but more so when used in combination. We illustrate this in the case studies that follow, but first outline the basic mechanics of the script on which our procedure is based.

When analyzing a modest number of records, manual inspection has been shown to be an effective technique (Iversen et al., 2007), but when dealing with "large-scale applications… it is necessary to automate the disambiguation process as much as possible, to keep the approach feasible and easy to maintain over time, as more and more data becomes available." (D'Angelo et al., 2011). Our approach is somewhat of a hybrid, relying initially on author identifiers, then commonalities among fielded data other than author identifiers, and finally manual inspection. To achieve name consolidation independent of author identifier matches, we have developed a procedure that is used with bibliometric software called VantagePoint.[8] While the application of our technique does not exclusively depend on VantagePoint, this is the software we find most efficient in the following analysis. The script we developed to implement this procedure is made available at the VantagePoint Institute (VPI).[9] It's designed to implement our name disambiguation procedure in a way that significantly reduces manual effort on the user's part. Those who seek to replicate our procedure independent of VantagePoint can do so by manually following the method we outline, but we note that the manual application of our procedure takes a significant amount of time and effort, especially when working with larger datasets.

Our script begins by asking the user for a surname and a first initial (for any author of interest). It then prompts the user to select a WOS field on which to consolidate author names. After this the user is prompted to point to the name of the authors field, and finally asked to identify a specific author name (referred to by the script as the primary author) within this field whom the user knows to be a true positive (a suggested approach is to point to an author name associated with one of the records that has the author's ORCID iD or email address attached to it). Figure 1 provides a visual:

The script proceeds to identify and combine all author names sharing the primary author's surname and first initial of his or her first name who share commonalities in the WOS field on which the user was prompted to consolidate author names. Initial dataset size is significantly reduced, and after the procedure is finished, the

[8] see www.thevantagepoint.com
[9] see http://vpinstitute.org/wordpress

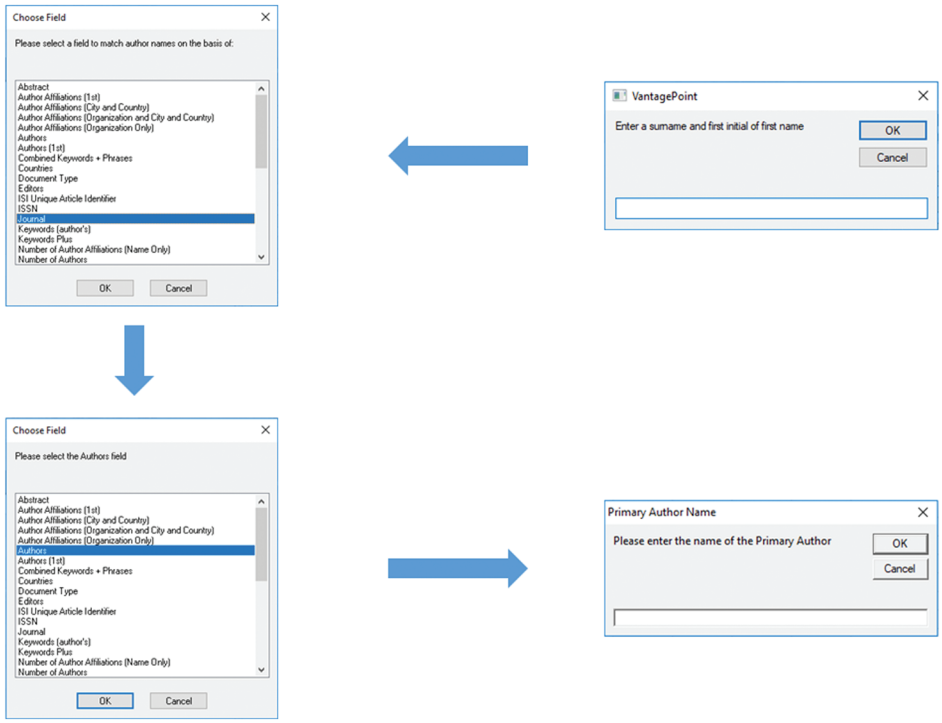

Figure 1. The progression of user inputs (starting in the top right and ending in the bottom right).

user is left with a much smaller (and more manageable) dataset to manually inspect (and/or apply additional disambiguation techniques to). Other studies (e.g., Song et al., 2015; Amancio, 2015) advocate using a combination of approaches, and doing so in conjunction with the procedure advanced in this paper is expected to produce results of both high precision and recall.

3.1 Case Study #1: Alan Porter (Georgia Tech and Search Technology)

Alan Porter has served as Professor Emeritus at the Georgia Institute of Technology and Director of R&D at Search Technology since 2002. As of late 2017 he has been published 234 times on WOS. We know his exact publication count after asking him to confirm publications belonging to him from a WOS CORE Collection search for Author = Porter, A (which resulted in 3,617 records). According to the U.S. Census Bureau, the surname Porter is the 159th most common in the United States. His publication count, along with the fact there are a number of additional authors with the same name as himself indexed on WOS, make him an ideal case study for present purposes. A WOS search for Professor Porter's work solely on the basis of

his ORCID iD (or his ResearcherID) yields only 40% of his scholarship on WOS, which points to the drawback of relying exclusively on author identifiers for name consolidation or bibliometric purposes (at least for some authors—it's noted that other authors have full and complete ORCID coverage, but without asking them to personally verify results, it's difficult to know what their author identifier coverage is).

If we apply the match technique advanced in this paper to the Source field (in the 3,617 records from a search for Author=Porter, A) our dataset moves from 3,617 records to 2,377 (a reduction of 34%). Proceeding in similar fashion for other match fields in our dataset we have the following:

Table 1. Match Results for Alan Porter.

Match field	Original dataset size	Reduced dataset size	Percent reduction
Source	3,617	2,377	34%
Co-authors	3,617	2,465	32%
Title	3,617	2,826	22%
ISSN	3,617	2,486	31%
Publication Year	3,617	2,905	20%
Affiliation	3,617	NA	NA
Cited References	3,617	NA	NA
Email	3,617	NA	NA
All of the above	3,617	1,750	52%

Fielded data for affiliation, cited references and email addresses was not available in significant enough quantity to be of use for purposes of our procedure, but among match fields that were available, we note that Source, Co-authors and ISSN all reduced the initial dataset by more than 30%. If we apply all of the above match fields to the initial dataset it is reduced by more than 50% and 233 out of Alan Porter's 234 true positives are retained (the one true positive that was lost occurred during the co-authorship round of reduction).

From our first case study, note that the more fielded data that is available and the higher the coverage for the same the better. Matching on the basis of email address would have been ideal, but that data was not available. Moreover, data for Cited References, a common match field for name disambiguation purposes, also was not available. Among the match fields that were available, Source, Co-authors and ISSN performed comparably (in terms of dataset reduction). We also note that while matching on the basis of individual fields reduces the initial dataset by as much as 34%, matching on the basis of multiple fields reduces the initial dataset by more than half (i.e., the more match fields used the better). The new dataset still contains a substantial degree of noise, but the user is now approaching a much more manageable dataset to manually reduce.

3.2 Case Study #2: Zhong Lin Wang (Georgia Tech)

Zhong Lin Wang serves as the Hightower Chair in Materials Science and Engineering at the Georgia Institute of Technology. He is a Regents' Professor at Georgia Tech and in 2009 was elected as a foreign member of the Chinese Academy of Sciences. A WOS CORE Collection search for AUTHOR=Wang, ZL from 2009 to the present yields 4,810 results. During this timeframe he has been published 700 times on WOS. We know his exact publication count after asking him to confirm publications from a list of WOS search results. According to the Chinese Ministry of Public Security, Wang is the most common surname in mainland China, making Professor Wang a unique challenge for name disambiguation purposes. Applying our match technique to a WOS search for him results in the following:

Table 2. Match Results for Zhong Lin Wang.

Match field	Original dataset size	Reduced dataset size	Percent reduction
Source	4,810	3,560	26%
Affiliation	4,810	4,147	14%
Web of Science Category	4,810	4,173	13%
Co-authors	4,810	4,175	13%
Title	4,810	3,794	21%
ISSN	4,810	3,555	26%
Publication Year	4,810	4,349	10%
Cited References	4,810	4,347	10%
Email	4,810	3,349	30%
All of the above	4,810	2,894	40%

Fielded data was in greater supply for this case study. As might be expected, the email address field performed the best of all match fields in terms of list reduction results. Source and ISSN tied for second place. As was the case with Porter, the more match fields used, the greater the total reduction. Applying our script for each of the above fields results in 40% list reduction, while retaining 695 (99%) true positives. The retention of all true positives is particularly difficult for scholarship as voluminous as Wang's given that the likelihood of misspelled, incomplete and/or mis-categorized records grows substantially.

3.3 Case Study #3: Haesun Park (Georgia Tech)

Haesun Park serves as Professor of Computational Science and Engineering at Georgia Tech. She is an Institute of Electrical and Electronics Engineers Fellow, as well as a Society for Industrial and Applied Mathematics Fellow. Professor Park has co-authored more than 100 articles in peer-reviewed journals and conferences. Given that Park is the 3rd most common surname in Korea,[®] disambiguating

[®] See https://en.wiktionary.org/wiki/Appendix:Korean_surnames

scholarship associated with her name also poses a challenge. Park's publications are identified with benefit of her personal website.[18]

A WOS CORE Collection search for AUTHOR=Park, H from 2010 to the present yields 23,298 results. We select this timeframe because the number of results grows considerably if no time constraints are used. Applying the match technique, Park's WOS scholarship results in the following:

Table 3. Match Results for Haesun Park.

Match field	Original dataset size	Reduced dataset size	Percent reduction
Co-authors	23,298	8,527	63%
Source	23,298	14,174	39%
Affiliation (Organization Only)	23,298	20,867	10%
Title	23,298	4,978	79%
ISSN	23,298	14,762	37%
1st Author	23,298	14,791	37%
ORCID iD	23,298	3,288	86%
Researcher ID	23,298	2,903	88%
Web of Science Category	23,298	21,906	6%
Publication Year	23,298	23,094	1%
Country	23,298	23,044	1%
All of the above	23,298	2,319	90%

Interestingly enough our procedure works particularly well when applied to the scholarship of Park, achieving a total list reduction of 90% (while retaining all true positives). A number of explanations might account for this. Park had excellent coverage for the match fields that appear in Table 3 (her mean coverage for these fields is 88%). Having a high coverage in more specific match fields (e.g., author identifier, email, co-authors, etc.) will produce more precise results and greater list reduction for the procedure we propose. In addition, our third case study had a more modest number of publications indexed on WOS. The sheer volume of research output by our first two case authors makes their list reduction more problematic. We note that volume of output on WOS is sensitive to a number of factors; one of which is the subject area in which the scholar under consideration publishes.

4 Results Compared

Comparing results from the preceding case studies yields the following:

[18] See https://www.cc.gatech.edu/~hpark/

Table 4. Comparison of Disambiguation Results for the Three Cases.

	Porter	Wang	Park
Top 3 list reduction match fields	Source (34%), Co-authors (32%) and ISSN (31%)	Email (30%), Source (26%), ISSN (26%)	Researcher ID (88%), ORCID (86%), Title (79%)
Original dataset size	3,617	4,810	23,298
Total list reduction	52%	40%	90%
Number of true positives lost	1	5	0

From Table 4 we note that Porter and Wang share Source and ISSN in their three most effective match fields, while the most effective match fields for Park were author identifiers and Title. Interestingly enough a large number of titles (274) were used in multiple records, which might be expected when a large number of document types are present (Park had 16 to be exact). We find that match field effectiveness is in large part a function of coverage. Comparing original dataset size, the timeframe analyzed for each case study is not the same, nor are the subject areas in which they publish. For reasons mentioned previously our procedure is more effective when applied to our third case study, both in terms of list reduction and 100% retention of true positives (case studies #1 and #2 had a 99% retention of true positives, so not bad).

5 Discussion

In this study we present a practical and replicable strategy for addressing the author uncertainty problem. Put another way, this paper uses a "forward stagewise" approach, which finds an optimal combination of variables that disambiguates author names by sequentially adding variables to the algorithm without adjusting the parameters of those that already produce some positive disambiguation results. In this paper we begin with telling match variables like author identifiers and email addresses and move on to (less precise but still useful) variables like co-authors, affiliations, journals, titles and ISSNs. An alternative is to use a "backward stagewise" approach that would begin with all the variables considered in the set and then sequentially remove them until disambiguation capability is significantly reduced. We did not use the latter approach because we found the former to yield more parsimonious combinations of fields, but future research might apply the latter and compare the approaches.

While machine learning is considered authoritative by many, we do not see it as practical or replicable. The procedure advanced in this paper is both practical, replicable and relatively user friendly. It might be categorized into a space between ORCID and machine learning. Machine learning approaches typically look for commonalities among citation data, which is not always available, clean, structured

or easy to work with. Coverage can be an issue. The procedure we advance is intended to be applied across numerous fields in a dataset of interest (e.g., emails, co-authors, affiliations, etc.), resulting in multiple rounds of reduction.[©] It provides a viable and useful alternative for situations where there are no reliable disambiguation systems in place (and/or situations where data such systems rely on is less than robust).

Results of our procedure indicate that effective match fields include author identifiers, emails, source titles, co-authors and ISSNs. While the script we present is not likely to result in a dataset consisting solely of true positives (at least for more common surnames), it does significantly reduce manual effort on the user's part. Dataset reduction (after our procedure is applied) is in large part a function of (a) field availability and (b) field coverage. No matter how large the initial size, dataset reduction is likely to be significant when field availability and coverage are robust. Even when these are not robust, however, the procedure we present provides a useful first pass (at dataset reduction)—one very likely to save the user time and manual effort. Other reduction techniques (manual investigation being just one) can then be applied to further reduce dataset size.

Results from our procedure are sensitive to a number of factors:

(i) *How common the surname under consideration is and if it changes*

The more common the surname under consideration the more challenging it will be to identify a specific author of interest. While Porter's initial dataset was reduced by 52%, Wang's dataset reduced by 40%. This finding is unsurprising given that Wang is a relatively more common surname than is Porter. While we're satisfied that a dataset reduction of 40% was achieved for the most common surname in the most populous country on the planet, we note that more common names are likely to involve greater manual effort after our procedure is applied if field availability and coverage are not robust.

In additional to commonality concerns, changes in surname (the result of a legal or matrimonial name change) likely will pose a challenge for our procedure. Tang and Walsh (2010) note that scholars who marry and have their family name changed further contribute to name variety challenges. Not everyone who becomes married has a name change, but when analyzing those that do, the procedure we advance (which takes family names as its starting point) will be hindered in its ability to match these names.

[©] It's noted that more aggressive matching can significantly boost the number of positive matches, but at the potential expense of incurring false negatives.

(ii) *Match field availability, coverage and homogeneity*

In the case of Porter, fielded data were not available for Affiliation, Cited References and Email Address. Had these data been available, it is likely that the reduction of his initial dataset would have been significantly more than 52%. Moreover, among fielded data that were available, coverage was sparse at times, rendering certain fields unusable for present purposes. We recommend a coverage threshold of 80% or greater to obtain valid results—i.e., it's recommended that 80% of the records in the user's dataset contain information for the match field (coverage lower than this can let true positives go undetected).

It should be noted that match data for scholars who have changed careers and/or fields dramatically is likely to behave differently. If the names 'Doe, J' and 'Doe, John' are associated with a dozen or more affiliations it's unlikely they're the same person, unless this individual has made multiple organization and/or career changes. While such is possible it typically isn't likely, and is expected to be associated with older scholars (who have had time to make multiple career adjustments). Results are expected to be less precise (i.e., we would expect a lower removal of false positives) for individuals with many and/or dramatic field or career changes.

(iii) *The subject area of the author of interest*

Wang publishes in subject areas typically associated with a high number of co-authors who enjoy greater publication counts than social science (and other) disciplines. The user ought to bear in mind that, as was the case with common surnames, subject areas with higher publication counts can involve more manual effort after our procedure is run. The user ought to keep in mind as well that when evaluating a reduced dataset, if 'Doe, John' and 'Doe, J' publish in two very different subject areas it is typically unlikely they are the same person. This isn't always necessarily the case, however. This is can be sensitive to the scholar's research interest(s) and/or the length of time he or she has been publishing for. For example, one of the case studies on this paper (Alan Porter) has been published more than 200 times on WOS over a span of more than 45 years. He's been published in more than 20 Web of Science Categories, and these intersect with the Social Sciences, Physical S&T, Computer Science and Engineering, Biology and Medicine. He has also concurrently served on two very different academic departments: (i) Industrial & Systems Engineering and (ii) Public Policy. We find a similar situation for another of our case studies (ZL Wang), who's been published 700 times in 41 Web of Science Categories. Hence, while a diversity of publications across multiple subject areas would likely indicate that 'Doe, J' and 'Doe, John' are not the same person (especially if they are younger and less prolific), this isn't always the case (especially for more established and interdisciplinary scholars).

(iv) *The ethnicity of the author of interest*

A search for Asian names is likely to be different than a search for Western names. The transliteration of differently spelled Asian names into the same English name, notable increase in the number of active scientists, and the sharing of a modest number of family names by a large number of scholars, all add to a number of factors that make Asian name disambiguation a true challenge (Tang & Walsh, 2010). When profiling the work of an Asian scholar using our procedure it's advisable to use as many and as precise match fields as possible. Other notable challenges not directly addressed here include multiple spellings in many Middle Eastern names (e.g., Mohammed) and more complicated Hispanic names (multiple parts, variations in ordering, etc.).

(v) *The country of the author of interest*

Our procedure uses author identifiers as a starting point. As is noted by Youtie and colleagues (2017), author identifier coverage varies significantly by country. Generally speaking, coverage tends to be stronger in Europe and weaker in Asia. When analyzing research from a specific country or set of countries using the procedure advanced here, the user should be cognizant of accompanying author identifier coverage. If author identifier coverage is poor or nonexistent for a given author, other match fields can be used, but we only advocate using them if they are reasonably precise and have adequate coverage (i.e., 80% or higher).

(vi) *The number and type of match fields used*

As mentioned, we take the position that matching on the basis of either author identifier or email address is adequately conclusive. Matching on the basis of other fields is less straightforward, however. As a general principle we find matches made on the basis of three or more of the following to be adequate: co-authors, affiliations, journals, titles and ISSNs. For fields less precise than these (e.g., subject areas, publication years, cited authors, cited journals, keywords, etc.), we advocate a higher threshold. After matches are made, the procedure advanced in this paper will produce match results for the user's consideration. If names A and B share the same family name and first initial, as well as commonalities among one or more key match fields (i.e., co-authors, affiliations, journals, titles or ISSNs), manual investigation is encouraged.

After all matches are made, a number of useful match fields will ideally be available for manual inspection. The subject areas field can be revealing, for instances—i.e., do the subject areas seem relatively homogeneous? Are there outliers? Are subject areas all over the place (which should raise red flags)? The publication year field is also revealing—are publication years close enough to one another to represent the work of a single author? Are there outliers? Do publication

years increase in steady progression or move forward in leaps and bounds? More precise fields, like author identifiers and email addresses, allow the user to weed out false positives more quickly. If two authors sharing the same surname and first initial have a different email address or author identifiers this is immediate cause for investigation. While it is conceivable that the same author has different email addresses (or author identifiers) on different publications, oftentimes a discrepancy in these categories signals the presence of false positives. In general, the higher the coverage threshold used when matching fields, and especially more precise fields, the less manual investigation will be required on the user's part.

Acknowledgements

The authors thank ZL Wang for his assistance with this study. This study was undertaken with support from the US National Science Foundation under Award 1645237 (EAGER: Using the ORCID and Emergence Scoring to Study Frontier Researchers).

Author Contributions

Stephen Carley (stephen.carley@searchtech.com) performed the analysis and conducted interviews. Alan Porter(aporter@searchtech.com) provided a case study and validated results. Stephen Carley, Alan Porter and Jan Youtie(jy5@mail.gatech.edu) wrote the first draft, and Stephen Carley revised the manuscript.

References

Amancio, D.R., Oliviera Jr., O.N., & Costa, L.D.F. (2015). Topological-collaborative approach for disambiguating authors' names in collaborative networks. Scientometrics, 102(1), 465–485.

Cota, R.G., Gonçalves, M.A., & Laender, A.H.F. (2007). A heuristic-based hierarchical clustering method for author name disambiguation in digital libraries. In Proceedings of the XXII Brazilian symposium on databases (pp. 20–34). João Pessoa, Paraiba, Brazil.

D'Angelo, C.A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. Journal of the American Society for Information Science & Technology, 62(2), 257–269. doi:10.1002/asi.21460

Diesner, J., & Kim, J. (2016). Distortive Effects of Initial-Based Name Disambiguation on Measurements of Large-Scale Co-authorship Networks. Journal of the Association for Information Science and Technology, 67(6), 1446–1461.

Ferreira, A., Goncalves, M.A., &Laender, A.H.F. (2012). A Brief Survey of Automatic Methods for Author Name Disambiguation. Sigmod Record, 41(2), 15–26.

Gasparyan A.Y., Nurmashev B., Yessirkepov M., Endovitskiy D.A., Voronov A.A., & Kitas G.D. (2017). Researcher and Author Profiles: Opportunities, Advantages, and Limitations. J Korean Med Sci., 32(11), 1749–1756. https://doi.org/10.3346/jkms.2017.32.11.1749.

Han, H., Giles, C.L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In Proceedings of the 4th ACM/IEEE-CS joint conference on digital libraries (pp. 296–305), Tuscon, USA.

Haesun Park (2006). Georgia Tech. Retrieved from https://www.cc.gatech.edu/~hpark.

Hussein, I. & Asghar, S (2017). A survey of author name disambiguation techniques: 2010–2016. The Knowledge Engineering Review. 32. 10.1017/S0269888917000182.

Iversen, E.J., Gulbrandsen, M., & Klitkou, A. (2007). A baseline for the impact of academic patenting legislation in Norway. Scientometrics, 70(2), 393–414.

Lang, F., Chavarro, D. & Liu, Y. (2016). Can Automatic Classification Help to Increase Accuracy in Data Collection? Journal of Data and Information Science, 1(3), 42–58.

Macroberts, M.H., & Macroberts, B.R. (1989). Problems of citation analysis: A critical review. Journal of the American Society for Information Science, 40, 342–349.

ORCID Connecting Research and Researchers (2010). ORCID. Retrieved from https://orcid.org.

ResearcherID (2008). Thompson Reuters. Retrieved from www.researcherid.com.

ResearcherID (2011). Clarivate Analytics. Retrieved from https://clarivate.com/products/researcherid.

Shin, D., Kim, T., Choi, J. & Kim, J. (2014). Author name disambiguation using a graph model with node splitting and merging based on bibliographic information. Scientometrics, 100(1), 15–50.

Smalheiser, N.R., & Torvik, V.I. (2009). Author name disambiguation. In B. Cronin (Ed.), Annual review of information science and technology (Vol. 43). Maryland, USA: American Society for Information Science and Technology (ASIST).

Song, Y., Huang, J., Councill, I.G., Li, J., & Giles, C.L. (2007). Efficient topic-based unsupervised name disambiguation. In Proceedings of the 7th ACM/IEEE joint conference on digital libraries (pp. 342–351). Vancouver, BC, Canada.

Song, M., Kim, E.H.J., & Kim, H.J. (2015). Exploring author name disambiguation on PubMed-scale. Journal of Infometrics, 9(4), 924–941.

Tang, L., & Walsh, J.P. (2010). Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. Scientometrics. doi: 10.1007/s11192-010-0196-6

Who uses Scopus (2015). Elsevier. Retrieved from https://www.elsevier.com/solutions/scopus/content.

Wiktionary (2007). Wikimedia Foundation. Retrieved from https://en.wiktionary.org/wiki/Appendix:Korean_surnames.

Youtie, J., Carley, S., Porter, A.L. & Shapira, P. (2017). Tracking researchers and their outputs: new insights from ORCIDs. Scientometrics, 113(1), 437–453.

Zhou, P. & Leydesdorff, L. (2006). The emergence of China as a leading nation in science. Research Policy, 35(1), 83–104.