Title: A Multi-Field Approach to the Author Uncertainty Problem

Stephen Carley stephen@iisco.com IISC
Alan Porter alan.porter@isye.gatech.edu Georgia Institute of Technology
Jan Youtie jy5@mail.gatech.edu Georgia Institute of Technology

The ability to identify individual scholars is fundamental to assessing productivity, mobility, collaboration and performance. In more recent times this task has become quite challenging. Techniques that rely on manual hand-checking are increasingly less viable in the age of Big Data. Despite the advent of author identifiers and name disambiguation algorithms, bibliometricians continue to be frustrated in attempting to disambiguate large-N common names. The rise of scientific powers like China, which produce a substantial volume of scholarship and have a large number of scholars sharing common last names, contribute to authorship uncertainty.

This study considers previous attempts to disambiguate large-N common names, along with drawbacks endemic to each. It goes on to advance its own approach to the authorship uncertainty problem. Results are obtained by a VantagePoint[1] script that proceeds via multiple rounds of reduction. Starting with a Web of Science (WOS) download of a common surname followed by a forename first initial, the script iteratively reduces an initial dataset by considering commonalities (or lack thereof) in fielded data shared by two names. If similar names belonging to separate records meet the script's commonalities threshold in Field A they then advance to a Field B comparison. If they satisfy the Field B threshold they proceed to a Field C comparison, and so forth. The order of field-based comparisons is nontrivial as commonalities in certain field matches are more telling than they are in others. After all reductions are made, a body of scholarship results that can be evaluated in terms of how closely it approximates the true publications for a given scholar. Results also allow for consideration of how the procedure advanced herein compares against alternate approaches.

The current study focuses attention on the surname 'Wang,' which happens to be the most common surname in mainland China, accounting for some 7.25% of all last names in the most populous country on the planet. As such, it provides a challenging case study. The authors (who are affiliated with Georgia Tech) are in personal contact with one Professor ZL Wang (Georgia Tech), who has authored more than 1,000 articles indexed on WOS. Identifying those true positives – i.e., papers that legitimately belong to the ZL Wang who works at Georgia Tech – can be likened to searching for approximately 1,000 needles in a sizeable haystack (an April 2017 WOS search for the author name "Wang, ZL" yields a whopping 7,279 results). Fortunately for us, after several in-person interviews we were able to obtain the solid list of true positives (from within the larger corpus of search results) for this author. We use this list as a benchmark against which to evaluate the performance of conventional approaches to the authorship uncertainty problem, as well as the approach advanced in this paper.

For the sake of convenience we restrict search results to those that appear from publication year 2009 onward. An April 2017 WOS search for the author name "Wang, ZL" yields 4,810 results (for publication years 2009-2017). Among this body of scholarship, 701 records (or 14.57%) are true positives for ZL Wang (Georgia Tech). The technique used in this study starts with the larger corpus of 4,810 records and proceeds to iteratively reduce it down to the true number (701) for the ZL Wang at Georgia Tech on the

[1] See https://www.thevantagepoint.com

basis of successive rounds of field comparisons among records containing similar names. In the context of this study two names are designated 'similar' if they share an identical surname and first initial. By way of illustration – because the names 'Wang, ZL' and 'Wang, Zhonglin' (which both appear in the corpus of 4,810 results) are similar, the records to which they belong are compared based on commonalities among the references they cite. If any such records share enough common cited references, they are then compared on the basis of the number of coauthors they have in common, and so forth. As will be seen, commonalities among certain fields (e.g. coauthors) are more telling (for purposes of name disambiguation) than they are in other fields. If records belonging to two similar names satisfy a threshold of such commonalties, we designate these as belonging to one and the same author. Comparisons are made at the level of records for which two similar names are attached (in lieu of the author name level) in light of the fact that in a number of instances two separate authors in our dataset share the exact same (first and last) name.

Once all matches are made the results of the procedure used in this study are assessed in relation to the true list of articles authored by ZL Wang (Georgia Tech). Results for more conventional approaches are also compared to the list of true positives for ZL Wang (Georgia Tech). A discussion of the efficacy and drawbacks of each follows.