Annex 2: Details of data set construction

June 24, 2015

1 Affiliation identification and histories

There are 536,454 author-article combinations included in our database, of which 31% lack affiliations We recover affiliation information for many of these authors by applying the procedures developed by Tang and Walsh (2010), as implemented in Agrawal et al. (2013). For each record without author's affiliation we check whether there is another record with the same author name (full surname and name or full surname and initials) with an affiliation. We assign this latter affiliation to the missing record as long as both articles cite, at least, two articles that are not highly cited. The low citation benchmark is set at less than 50 citations. This increases the author-article combinations with affiliation information for some authors from 69% to 80%. Of those, 84% have affiliations for all authors.

We impute affiliation information for years in which an author does not publish by using his or her affiliation before or after those years. Our algorithm uses, iteratively, the closest information relative to the information gap. For example, suppose that author A published an article in 1990 when she was affiliated to MIT, and then published her next article in 1994 when she was affiliated to Princeton. In this example, we have holes in the affiliation history of this mathematician from 1991 to 1993. In the first iteration, the algorithm will fill the 1991 hole with information from 1990 (the closest available year), and the 1993 hole with information from 1994. After the first iteration we will still have a hole for the year 1992. We apply the second iteration to the algorithm. In this case, the author will have a double affiliation for the year 1992, because she has two different affiliations in the closest years (1991 and 1993).

2 Self-citation

To identify self-citations, we developed a unique author code that combines data from WOS, MGP and zbMATH databases (see below). MGP and zbMATH provide the name and surname of the authors, plus a unique author identification code. WOS only provides the surname and initials of the author. As zbMATH identifies the author at the article level, for those articles included in the zbMATH database, we were able to match WOS authors with zbMATH author codes. The personnel at zbMATH also provided us with a correspondence between zbMATH author codes and MGP author codes. For the rest of authors, we assigned a zbMATH author code if there was only one author code for a surname+initials combination. For the remaining cases, we created a unique author code. To be conservative, we consider a self-citation if any of the citing author has the

same zbMATH code as any of the cited authors; and when any citing author has the same surname and initials of any of the cited authors.

References

Agrawal, A., McHale, J., and Oettl, A. (2013). Collaboration, stars, and the changing organization of science: Evidence from evolutionary biology. NBER Working Papers 19653, National Bureau of Economic Research, Inc.

Tang, L. and Walsh, J. P. (2010). Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, 84(3):763–784.