Draft version October 12, 2016
Preprint typeset using LaTeX style AASTeX6 v. 1.0

, , ,

PUBLISHING LIFETIMES OF AMERICAN ASTRONOMY PHDS: A POST-2008 COLLAPSE

Peter Yoachim[1], Staša Milojevic[1], xxx-author 2, and xxx-author 3

[1]Department of Astronomy, University of Washington, Box 351580, Seattle WA, 98195; yoachim@uw.edu

ABSTRACT

XXX-blah blah blah, abstract

Keywords: Employment, Physics Education

1. INTRODUCTION

XXX-recast as exploring general phenomenon with ASt as a test discipline that has good bibliographic data.

It is difficult to determine the state of the Astronomy job market. Analysis of the science job market often lumps all sciences together, in which case Biologists dominate the statistics, or, if the sciences are separated, Astronomy is usually grouped with Physics (e.g., the Bureau of Labor and Statistics lists 20,000 Physicists and Astronomers jobs in 2014 [1]), where the number of Physicists outnumbers the Astronomers. The American Astronomical Society currently lists it's membership at around 7,000.

XXX-move to discussion. Seth et al. (2009) builds on the work of Metcalfe (2008) and looks at the rise of the postdoc era in Astronomy and Astrophysics. They show that as the inflation-adjusted US-Astronomy budget doubled over 20 years, there was a modest increase in PhD production and a huge rise in the number of temporary postdoc positions.

In this paper, we use the SAO/NASA Astrophysics Data System (ADS) to explore the ramifications of that growing postdoc bubble. While the NSF regularly surveys PhD recipients (e.g., http://nsf.gov/statistics/nsf07305/), even with high participation rates, surveys will tend to be biased against those that leave the field since they will be harder to contact. Thus, we propose to use the publication record of PhDs to track the overall health of the field.

Unfortunately, some faculty still rely on anecdotally information to claim, "Yet the [astronomy faculty job] market is no worse or better than it is has been for at least a decade or two." (via Charfman[2]). The goal of this paper is to check the veracity of this statement and quantify the astronomy job market by tracking the digital footprints Astronomers leave in ADS.

The ADS is an abstracting service offering a search interface into the scientific and technical literature covering astronomy, planetary science, physics, and the arXiv e-prints. ADS includes more than 12.2 million listings for peer-reviewed papers as well as conference proceedings, conference posters, grants and, importantly for this paper, PhD Theses.

Example of current tracking here [3]–note that they only get info on 48% of physics PhD recipients. We expect that ADS is much more complete in listing Astronomy PhD's than 48%. XXX-grab list of recent Prize fellows and see how complete the ADS listings are.

2. DATABASE CONSTRUCTION

All the code used for this paper is available on github[4]

We use the excellent *Python Module to Interact with NASA's ADS that Doesn't Suck*[5] to access records from the SAO/NASA Astrophysics Data System.

Our procedure for building a database of Astronomy and Astrophysics PhDs is as follows (throughout this paper, we use "Astronomy" and "Astrophysics" interchangeably). We select all PhD Thesis records for a given year and screen to include only those records that have an affiliation in the United States. We automate this process by checking the institution field to see if it includes the name of a US state, or includes one of 62 strings we manually compiled (e.g., "Harvard", "Portland", "Howard", etc.). The full list of strings is available with the source code.

For each US PhD, we then query ADS for all entries with the PhD author's last name and first initial with a date limit of 7 years pre-PhD to the present (we searched ADS in January, 2016). XXX-justify 7 years. We found

[1] http://www.bls.gov/ooh/life-physical-and-social-science/physicists-and-astronomers.htm

[2] http://jjcharfman.tumblr.com/

[3] https://www.aip.org/sites/default/files/statistics/employment/phd1yrlater-p-14.pdf

[4] https://github.com/yoachim/AstroHireNetwork

[5] https://github.com/andycasey/ads

8820 PhDs from 1997 to 2013, 2903 were determined to be unique PhD listings of astronomers granted from USA institutions.

XXX-discuss name disambiguation earlier.

For common names (e.g., "Williams, B"), our subsequent author query returns entries from multiple authors. To try and restrict the results to only those papers with the same author as the PhD thesis, we construct a network graph of the ADS results using Networkx Hagberg et al. (2008).

There can be variations in how author names are spelled, including capitalization variations (e.g., "VanderPlas" (VanderPlas et al. 2012) and "Vanderplas" (Vanderplas 2012)) as well as variation in the presence or absence of accent characters. To ensure we match authors correctly, we reduce all author name searches to the format "first initial, last name", and consider names to match if they have a Levenshtein distance less than three.

XXX-note that this is a new name disambiguation scheme. Cite "Accuracy of simple, initials-based methods for author disambiguation" (http://arxiv.org/abs/1308.0749)

Every ADS entry is a node in an author network. We draw an edge between nodes if they meet any of the following criteria:

- If the affiliations (of the possible thesis author) match, and the entries have publication years within 2 years or each other.
- If the entries have 3 or more authors in common in their author lists.
- If there are two authors and they are identical on the two entries.
- If the information contained in the abstracts are similar enough. We use sklearn Pedregosa et al. (2011) to convert the abstract text of each entry (if available) to a matrix of TF-IDF features. If two abstracts in the matrix exceed a threshold (ratio > 0.5), we consider them similar enough that the entries should be connected
- If the information contained in the titles of the entries are similar enough. As with the abstracts, we use a ratio of > 0.5. XXX–should show an example of this.
- If the relevant author's name matches exactly (identical spelling, capitalization, accent charters, middle initials/name, etc.), and the entries share two or more keywords.

For our affiliation matching, we consider the affiliations to match if one affiliation is contained in the other. For example, "University of Washington" is considered a match to "Department of Astronomy, University of Washington, Box 351580" but not "Washington University". We also compare the two affiliation strings with the python difflib SequenceMatcher, and consider any affiliations that have a similarity ratio greater than 0.7 as matching. For example "Berkeley" and "UC Berkeley" have a ratio of 0.8, and would thus be considered matching.

XXX-Run the name disambiguation on about 10 known names and say how well it performs.

This is an attempt at codifying the common-sense process one naturally takes to infer if it is the same author on two papers. With authors being able to identify their papers with services such as ORCID[6], it should soon be possible to build a network of papers known to be linked to a single author as a training set for machine learning algorithms.

As the final step, we select only the entries which are linked to the PhD thesis of interest. Some example ADS entry network graphs are shown in Figure 1.

Once we have selected only those entries we believe are connected to the thesis author, we demand that at least one of the entries be in a peer-reviewed astronomy journal to eliminate Physics thesis entries. XXX-note that it's all entries, not just first author. The list of publications we use to label someone as an astronomer are listed in Table 1. We consider an ADS listed publication to be a match if any of our selected journals are contained in the publication string, e.g., "Astrophysical Journal Letters" would count as a match because it contains "Astronomical Journal" which is in our list. This step is necessary to remove PhD authors who would be better classified as Physicists, Geophysicists, Planetary Scientists, etc. XXX-Do a manual test to make sure astronomy classification is working (I think all the Hubble Fellows should be AST, so that might be an easy check).

We do not limit our search to only first-author publications for either the publication graph construction or determining if the author meets our criterion for being labeled an "astronomer". We also do not limit the query to peer-reviewed journals, thus our query results include things like conference posters and grant proposals. We treat these results identically to journal articles when constructing the author networks.

The ADS publication must match at least one of:

[6] http://orcid.org/

Table 1. Publications Used to Classify Authors As Astronomers

Journal	abbr.
Astronomy and Astrophysics	A&A
Astrophysical Journal	AJ
Astronomical Journal	ApJ
Monthly Notices of the Royal Astronomical Society	MNRAS
Publications of the Astronomical Society of the Pacific	PASP

Thus, our definition of a US PhD astronomer is effectively a person who received a PhD from an institution in the United States and has published at least one paper in a major peer-reviewed English-language astronomy journal. XXX-move this up earlier

We record the thesis author's most recent ADS entry date (need not be a peer-reviewed paper) and a bunch of other stuff.XXX-clarify or delete.

For each US astronomer, we query ADS to look for any other PhD articles with the same first initial and last name (from any year). We flag the author as having a potentially unique name if this query only returns their thesis. The subset of potentially unique PhD names can be used to test our name disambiguation techniques.

One known issue is that ADS does not necessarily contain all Astronomy PhD dissertations. When we ran our code Pagnotta (2012) was not listed in ADS (but now appears). XXX-put this earlier, note that it's not expected to be a large effect.

To test our author linking algorithm, we take the list of Hubble Fellows[7] and manually trim the list down to those from US institutions. We expect all of these individuals should be in our ADS database. We find we can match 147 of 168 Hubble Fellows (87.5%). The 21 Hubble Fellows who failed to match span all the relevant years, and failed for a variety of reasons. For example, ADS lists Jose Preito's PhD as Katunaric (2009). Other authors fail to be identified as astronomers, or fail to link to their full publication network. However, we find it promising that we appear to correctly recover 87% of known astronomers with no particular biases (e.g., we are not biased against certain years or institutions).

XXX-discuss the few unlinked entries in the networks of my and Eric's networks

XXX–maybe put this in an appendix and just have a brief discussion on why we trust the name disambiguation.

Possible ways we could fail to recognize an author has published for as long as they actually have (i.e., networks are under-linked) If an Astronomer:

- simultaneously changes institution and area of study
- leaves the field for an extended time and re-joins in a different field
- changes their name (although authors can notify ADS of name changes)

Ways we could mistakenly claim an author has published for longer than they actually have, we construct networks which suffer from over-linking, if two authors have similar names and:

- overlap at the same institution
- similarly named co-authors

Thus there are two types of errors that can skew our results, over-linking our networks, connecting ADS entries that are not actually by the same author and under-linking, where we fail to recognize entries that are by the same author. To look for the effects of over- and under-linking, we flag authors where the ADS database only lists one PhD thesis as matching their last name, and first initial. These represent less-common names that should suffer much lower rates of over-linking. Figure 1 panels (a) and (b) show authors with unique names. These two network clouds suffer from no over-linking and a very low level of under-linking that does not hinder our ability to correctly measure the last year the authors had a first-author entry. Figure 2 shows an example network cloud where the under-linking is severe enough to make us incorrectly assume the author is no longer active as a first author.

Using the less-common name subset, we can also test for under-linking by assuming the names are unique and all the ADS entries returned for a name search should be linked. We find little evidence of under-linking. Comparing the less-common name results changes the retention rate by only $\sim$ 4%. When we test for under-linking, we find recent PhDs could be suffering from under-linking at the 5-8% level, while older PhDs suffer under-linking at lower rates (2-4%). This should be slightly expected, as younger astronomers have typically change institutions more recently, and many will have not yet published papers that link back to their previously established paper network cloud. The important point is that the relative over- and under-linking between PhD cohorts is similar. Thus, the recent dramatic (20%) drop we observe in astronomer retention is real and not an artifact of how we assign ADS entries to authors. XXX-odd place to spill the main point of the paper.

[7] http://www.stsci.edu/institute/smo/fellowships/hubble/fellows-list/

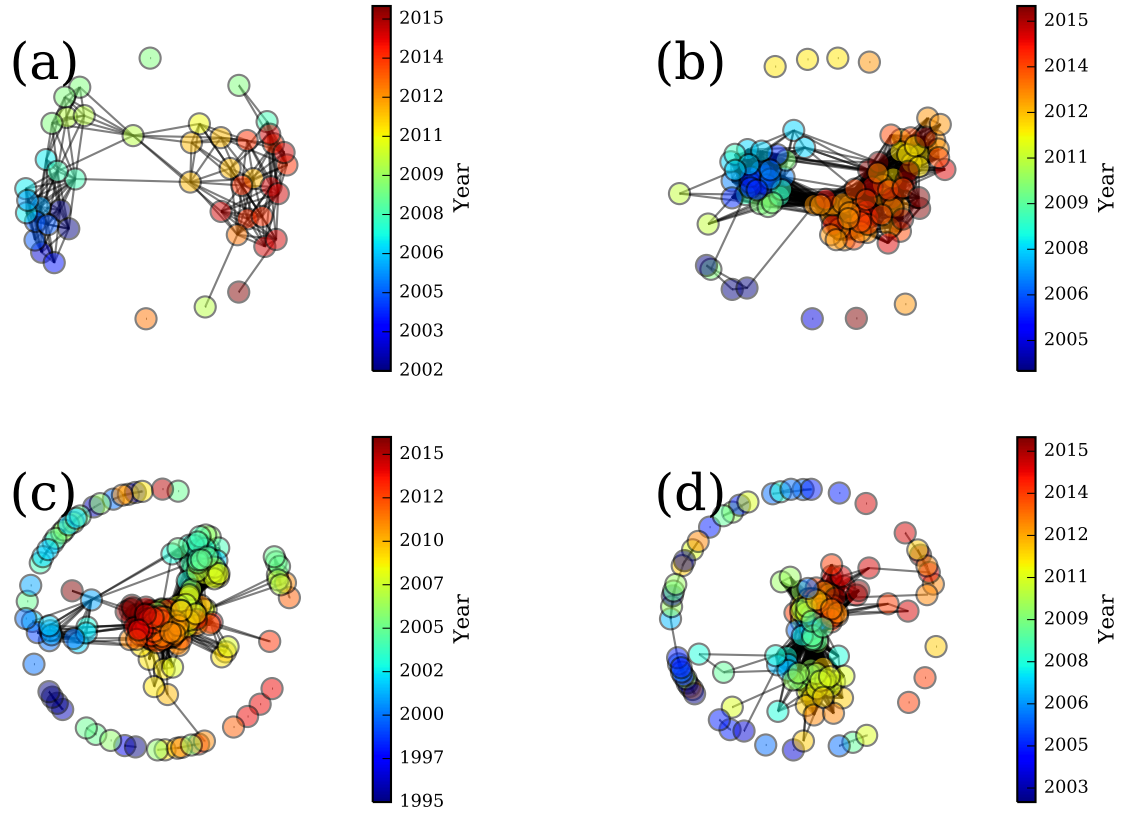

Figure 1. Examples of network graphs constructed to find papers linked to individual PhD thesis entries in ADS. (a) Network of ADS entries with the same author as Yoachim (2007) (46 entries, 44 linked to the PhD), (b) Network for Bellm (2011) (106 entries, 99 linked),(c) Network for Williams (2002), (315 papers, 270 linked) (d) Network for Williams (2010) (158 papers, 113 linked). Note, none of the linked papers for the two "Williams, B" PhDs overlap, suggesting our network construction procedure has correctly disambiguated the two authors.

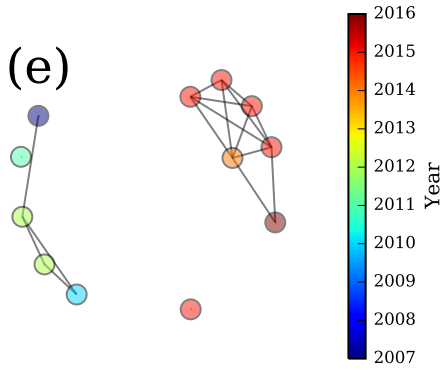

Figure 2. Example of an ADS networks that suffers from under-linking. (e) The under-linked network of Capelo (2012). A single ADS entry in the future could link a network back together.

It is worth emphasizing that while there are undoubtedly some cases where we have not accurately recorded an astronomer's publication network, we do not suffer the selection biases that go with survey-based studies and that our errors should be similar across PhD cohorts. Thus we should be able to draw strong conclusions about the relative differences between PhD classes.

3. RESULTS

XXX-introduce retention curves (and say why this is better than the silly replacement factor that I think I cite later.)

To test how much we are affected by over-linking ADS entries, we compare the retention curves of all astronomers to those with unique PhD names. xxx-should also then take the unique names and see how it changes if I use the last entry rather than the last linked entry.

Figure 4 is particularly striking, showing less than 50% of the 2009 to 2012 cohorts are still appearing as first authors on ADS entries. Note, these are any entries, we have not limited it to peer-reviewed papers.

Considering most PhD programs now take 6 years to complete, this implies that most post-2008 USA-PhD astronomers spend more time in graduate school than they do leading and publishing their own research projects post-graduation. Imagine if a corporation had a training program that lasted 3 months, and most of the employees who completed it left after 3 months post-training. This seems an odd place in the career path to place a

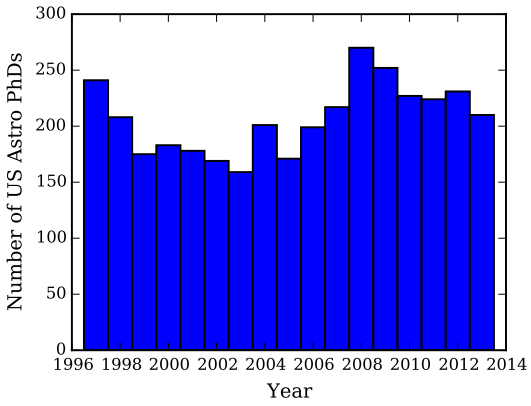

Figure 3. Number of US Astronomy PhD thesis entries found in ADS per year. XXX-compare to "official" stats from other sources.

bottleneck. xxx-move to discussion.

xxx-compare to using just peer-reviewed or author in any position.

There is a potential bias if authors publish as first authors on intervals longer than a year. If that is the case, we would expect the last few points of each cohort curve in Figure 4 to be slightly depressed.

Currently, 30% of PhDs are immediately leaving the field, double the fraction typical pre-2008.

4. DISCUSSION

The Astronomy labor market is not suffering from a funding crisis, but rather a funding *allocation* crisis. A simple thought experiment can help illustrate the problem. What would happen if the NSF was able to double it's Astronomy budget? Much of that money would be distributed as grants, that would then be used to hire graduate students and post docs. Very little of the money would go to funding permanent positions. Increasing, federal astronomy funding actually exacerbates PhD over-production.

With a large increase in funding from XXXX to XXX, driven largely by NASA, there has been an increase in Astronomy PhD production with little to no increase in long-term positions for these PhDs. This needs to be recognized as a failure of Astronomy funding agencies, as they paid to develop a talented workforce with no plans to retain that talent. Similarly, we must hold professional organizations responsible for failing to advocate for a funding stream allocated for the long-term health of the field. XXX-tone this down a bit to make it more neutral.

XXX-discuss drivers for PhD over-production. Some of it is department need for TAs, but I think that some extra is due to the funding increase, as seen in Fig 3.

The failure to retain early-career Astronomers is analogous to building 4 new telescopes, when one knows there will only be enough funding to operate 2. Such a lack of planning and wasteful use of research funding would not be tolerated with hardware, and it should not be tolerated with meatware. Given the stringent requirements to gain entry to an astrophysics PhD program, it seems unlikely that the community is successfully selecting the top potential early-career scientists to give long-term jobs to. Instead, we have created a large pool of talented PhDs where luck is the dominant factor in success rather than some sort of meritocracy. XXX-I don't really have support for this statement XXX-There is the arguemnt that students are cheap, so this could be producing the most science bang for the buck.

Cooray et al. (2015) title their paper "Astronomy job crisis". They contend that the solution to PhD overproduction is to change the culture within astronomy so that careers outside academia are the top priority of PhD students. This seems like a rather bizaar proposition that has no chance of succeeding. If one's goal is to peruse a career in industry, why would one start a PhD program in astronomy? While there is growing consensus that the astronomy community must be more supportive of those who chose to leave the field, we need to acknowledge the basic fact that students start astronomy programs because they are interested in having astronomy careers. When a recent top physics major declares they would like to study astrophysics, it seems fairly patronizing to tell them they should instead follow a course of study to prepare them for an industry job. Students will simply chose the programs that provide the training they are seeking. XXX-may not belong in research paper.

Cooray et al. (2015) never seem to contemplate the possibility of simply funding fewer PhD students. They claim, "It is hard for universities to limit the number of students they enroll...". On the contrary, universities hire as many grad students as they can afford to, and rarely extend beyond that limit lest they get a reputation for not supporting students.

Given the sharp recent change in PhD retention seen in Figure 4, it is important to also consider the resulting impact the change will have on recruiting new graduate students. We never see the students who chose to not apply for graduate school. This is an important observational bias. If astronomy PhD career prospects start to resemble Art History PhD career prospects, we should expect the technical skills of the two groups of graduate students to become more similar. With poor career prospects, astronomy risks losing the ability to recruit top students. Law schools recently saw a drop in top student applications with the collapse of the legal job

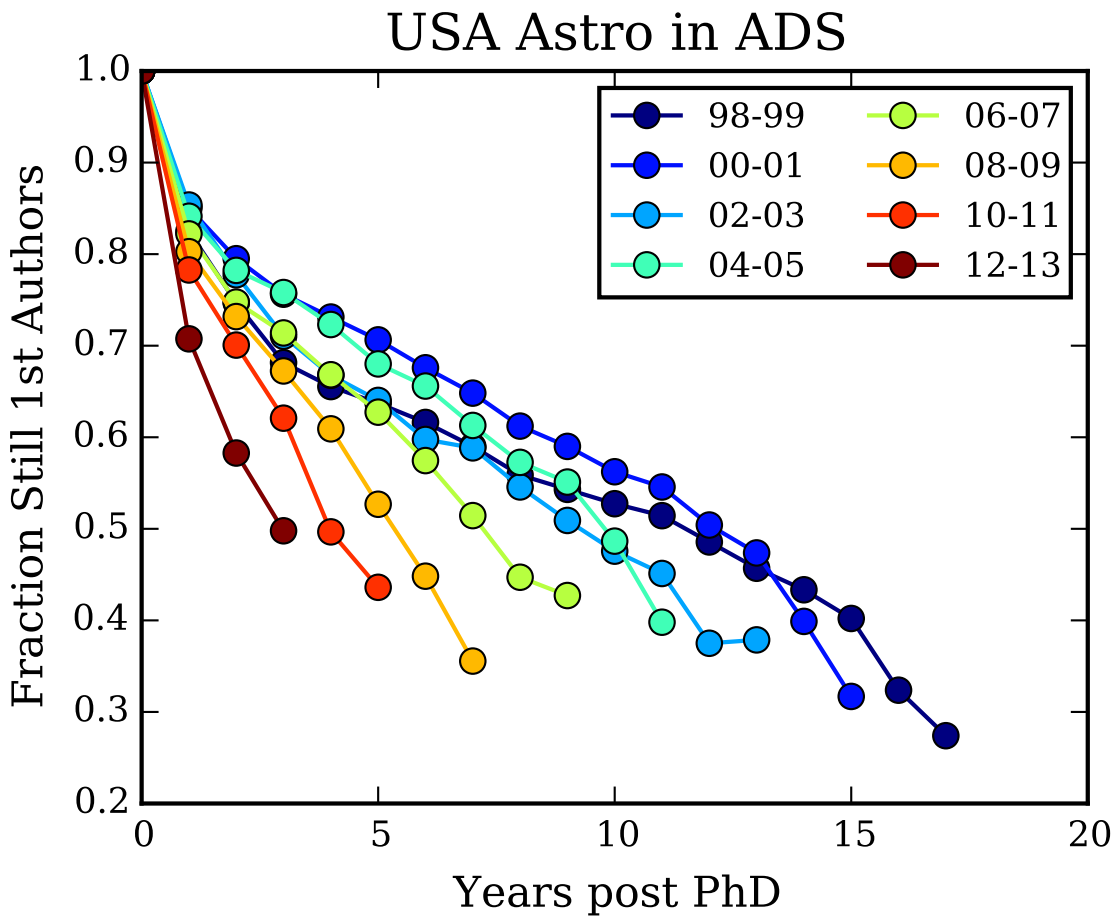

Figure 4. The fraction of Astronomy PhDs still active as first authors on any type of ADS entry (e.g., peer-reviewed journal articles, conference proceedings, grant proposals, arXive papers, etc). Error bars show ranges computed by comparing the curves to those of authors with unique names and unique names where all records are assumed to be linked.

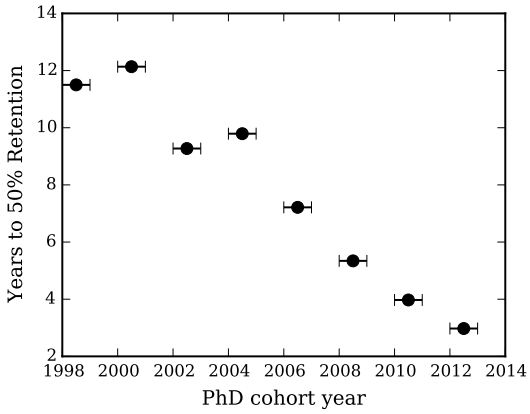

Figure 5. The same data as in Fig 4, but now showing the time it has taken for each PhD cohort to drop to 50% active. It appears we are losing young astronomers four times faster than in 2000.

market[8].

XXX-Point out that while the PhD retention has radically cahnged, it's hard to say what effect this has had on the field. We can't tell from our data if potential top astronomers are turning away from the field, or if there would be higher quality research done if more PhDs were retained and publishing longer.

There seems to be little hope that university departments will undertake a grass-roots effort to improve the career pipeline. Thus, it seems to fall to the funding agencies to enforce better behavior. One approach could be to require a minimum ratio of permanent-to-temporary PhDs (and PhDs in training) in a department before that department can receive federal funding. This prevents research universities from running degree-mill type programs, or from relying too heavily on adjuncts

[8] http://www.theatlantic.com/business/archive/2012/04/the-wrong-people-have-stopped-applying-to-law-school/255685/

for teaching. This also encourages states to fund colleges at a reasonable level to maintain access to federal funding. Another common proposed solution is to move federal funding away from university departments and more towards permanent positions at national laboratories and observatories.

4.1. *Future Work*

The ability to identify individuals via their PhD and track their career progress via their appearances in ADS entries opens up a number of possible future studies including

- Extending this analysis to countries beyond the US.
- For many author networks, it should be possible to programatically guess the author's gender, allowing one to compare career trajectories based on gender.
- One could analyze the network of PhD granting institutions and final hiring institutions to rank the prestige of graduate programs as in Clauset et al. (2015).
- The technique of linking publication histories to PhDs can easily be extended to other fields.

5. CONCLUSIONS

Using the bibliographic information in ADS, we find that retention of new American Astronomy PhDs is at a 15+ year low and has been gradually declining.

Some work for this paper was done at the 2016 American Astronomical Association Hack Day. Thanks to sponsors of AAS Hack Day 2016, The Large Synoptic Survey Telescope and Northrop Grumman.

This research has made use of NASA's Astrophysics Data System. Very heavy use. Thanks to Roman Chyla and the ADS staff who increased my daily query limit and gave helpful tips on optimizing ADS queries.

Software: Numpy (Walt et al. 2011)

Software: Matplotlib (Hunter 2007)

Software: Networkx (Hagberg et al. 2008)

Software: Pandas (McKinney 2010)

Software: A Python Module to Interact with NASA's ADS that Doesn't Suck™ https://github.com/andycasey/ads

Software: Scipy (Jones et al. 2001–)

Software: Sklearn (Pedregosa et al. 2011)

Software: Leven https://github.com/semanticize/leven

Software: Matplotlib_pubplots https://github.com/yoachim/matplotlib_pubplots

REFERENCES

Bellm, E. C. 2011, PhD thesis, University of California, Berkeley
Capelo, P. R. 2012, PhD thesis, Yale University
Clauset, A., Arbesman, S., & Larremore, D. B. 2015, Science Advances, 1
Cooray, A., Abate, A., Häußler, B., Trump, J. R., & Williams, C. C. 2015, ArXiv e-prints
Hagberg, A. A., Schult, D. A., & Swart, P. J. 2008, in Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA USA, 11–15
Hunter, J. D. 2007, Computing in Science & Engineering, 9, 90
Jones, E., Oliphant, T., Peterson, P., et al. 2001–, SciPy: Open source scientific tools for Python, [Online; accessed 2016-08-24]
Katunaric, J. L. P. 2009, PhD thesis, The Ohio State University
McKinney, W. 2010, in Proceedings of the 9th Python in Science Conference, ed. S. van der Walt & J. Millman, 51 – 56
Metcalfe, T. S. 2008, PASP, 120, 229
Pagnotta, A. 2012, PhD thesis, Louisiana State University
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. 2011, Journal of Machine Learning Research, 12, 2825
Seth, A., Agüeros, M., Covey, K., Danner, R., Jang-Condell, H., Metcalfe, T., Modjaz, M., Rasio, F., Redfield, S., Sheth, K., Waller, W., West, A., & Yoachim, P. 2009, Astronomy, 2010, 51P
Vanderplas, J. T. 2012, PhD thesis, University of Washington
VanderPlas, J. T., Connolly, A. J., Jain, B., & Jarvis, M. 2012, ApJ, 744, 180
Walt, S. v. d., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science & Engineering, 13, 22
Williams, B. F. 2002, PhD thesis, University of Washington
Williams, B. J. 2010, PhD thesis, North Carolina State University
Yoachim, P. 2007, PhD thesis, University of Washington