Stat 286 Final Project Report: Text Matching based on Document Embeddings

Kojin Oshiba, Wenshuo Wang, Han Yan

December 6, 2017

1 Introduction

There is a growing number of literature on text-based causal inference. The literature can be largely split into using text as (1) covariates , (2) treatments and (3) outcomes. In this paper we will focus on text as covariates. Specifically, we propose a method for matching that uses document embeddings as covariates. Document embeddings is a method that are shown to perform well in obtaining not only the topics but also the semantics and paragraph/sentence structures of documents. The key advantage of this approach is that we can match documents on document structures as well as latent topics and other features that are typically used in matching.

In particular, we apply document embeddings to measure gender bias in scientific community by comparing differences in citation counts of publications by male and female researchers. We define the outcome Y to be the citation counts and we use the texts as an important covariate X. To ensure unconfoundedness assumption holds, and the treatment variable does not affect other covariates, e.g. full text article and number of authors etc, that we are going to use for matching, we define our treatment T - "gender" - as reader perceived gender of the first author of the publication by reading the byline. Thus we are trying to estimate the casual effect of reader perceived gender of the first author of a publication on the number of citations received by the publication. We let $T = 1$ if readers perceive the first author as a female, i.e. the first author has a female given name, and $T = 0$ if it has a male given name.

Many studies have shown evidence that women are underrepresented in various academic areas. *Moss-Racusin et al. 2012*[1] conducted a randomized double-blind study to investigate the extent of bias in science faculties against female students. The experimental results show that "faculty participants rated the male applicant as significantly more competent and hireable than the (identical) female applicant". *Maliniak, Powers and Walter 2013* [2] found a systematic difference in citation counts between women and men in the international relations literature. They identify key attributes of an article, e.g. methodology, episteology, paradigm etc, treat them as covariates and apply a negative binomial model to estimate citation counts. *Caplar, Tacchella and Birrer 2016* [3] also found a citation difference between female first-authored and male first-authored papers published in several astronomy journals. They used male first-authored papers as training data to train the random forest algorithm, and then estimate the expected citation counts for female first-authored papers, and compare with the actual citation counts received.

2 Data

Due to data availability, and time constraints, we will specifically focus on papers published in the astronomy domain. We got most of the data from SAO/NASA Astrophysics Data System (ADS) Astrophysics Data System API (http://ads.harvard.edu/, https://github.com/adsabs/adsabs-dev-api). Specifically, we obtained astronomy papers according to the following conditions:

- Published between 2010 and 2012
- Published in either of the following journals: "Astrophysical journal" (ApJ), "Monthly Notices of Royal Astronomical Society" (MNRAS), "Astronomy & Astrophysics" (AA), "Nature" (NAT), "Science" (SCI)
- The following features were extracted: abstract, author, first author, citation count, keyword, publication date, title

We choose these five journals as they "encompassing the vast part of astronomical research today. Furthermore, they are well established journals with long historical records." [3] To obtain the full text of the papers, we used arXiv API (https://arxiv.org/help/api/index) to obtain the PDF urls. Then we downloaded PDF files from those urls, and parsed the PDF using an open source PDF parser (https://github.com/euske/pdfminer) to generate the raw texts. In the process, we removed data whose authors were not available or whose PDFs were not available on arXiv.

Once we obtained the full features, we estimated the genders of all the first authors. Since our treatment is defined as the reader-perceived gender of the first author, the gender of the first author of each paper is estimated using their given names. We use a Mathematica build-in classifier function to perform the task. We discarded units those gender can not be accurately identified, i.e. estimation accuracy is below 0.9. For example, usually it can be very hard for people to know the gender of a non-English first name after Romanization. In those cases, the treatment is undefined. This procedure also means we would treat a male author with female first name as female, because that's how the readers perceive his gender. In the end, we get a total of 893 samples, of which 238 papers' first author is perceived as female, and the rest 655 papers have male as the perceived first author gender.

3 Document Embeddings for Causal Inference

3.1 Doc2Vec

Document embeddings are gaining popularity in computer science literatures, and are high performing methods in sentence/document similarity measurement tasks. The basic concept of document embeddings is to have some model (often times, neural network) to convert a document into a vector representation. We will use one of the most widely used document embedding methods, doc2vec *Quoc Le et al. 2014*[5]. There are other popular document embeddings instances such as glove *Jeffrey Pennington et al. 2014*[6] but the methodology for causal inference is the same regardless of the embedding methods. Doc2Vec is an extension of a word embedding method word2vec which maps words to vector representations. Word2Vec is a three layer neural network whose task is to predict a center of a word in a subsentence from surrounding words. For example, if the sentence is "The quick brown fox jumps over the lazy dog", the training data generated is as follows:

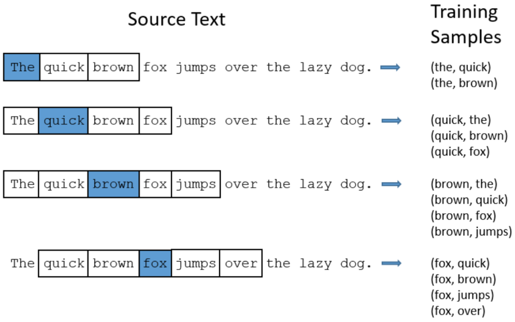

Figure 1: Demonstration for Word2Vec[7]

In the above diagram, the goal of the model is to predict the center word (colored in blue). The whole word2vec model looks like the following:

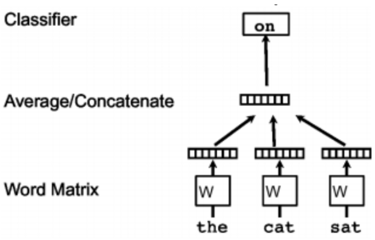

Figure 2: Model flow for Word2Vec[5]

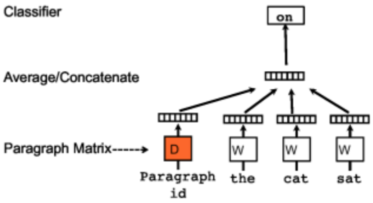

Figure 3: Model flow for Word2Vec[5]

Once the model is trained, the word vectors can easily be obtained from the hidden weight matrix, which functions as a look up table for the vector representations of each word.

Now, doc2vec can be understood as a simple extension of word2vec where, in addition to surrounding words, we add document ID's as additional covariates. That way, the hidden layer will not only learn the word embeddings but also document embeddings.

3.2 Advantages of Using Document Embeddings

Commonly, matching on text has been done using latent topic vectors obtained from models like STM (Structural Topic Models). STM has been widely used because it allows us to reduce the dimension of text covariates that are by default high dimensional. [8]

Using Doc2Vec over topic models like STM has an advantage of being able to capture document characteristics that are not just texts. In many cases, features that are not topics can highly influence the outcomes. In our example of measuring gender bias in science publications, it might be the case that papers of different lengths, certain logical structures, certain word usages might affect whether other scholar cite their papers. In such scenario, matching based on STM is inadequate. Doc2Vec offers a more holistic way of capturing the innate characteristics of documents.

Note that this does not mean that researchers need to choose between STM and document embedding methods. It is always possible to concatenate the covariates generated by the two methods and apply dimension reduction methods (e.g. PCA) if necessary.

4 Matching Methods

After summarizing every paper's full text into a vector using doc2vec method, we match every unit in the treatment group with a unit from the control group based on cosine similarity. This one-to-one match is done without replacement to ensure that matched pairs are independent of each other. Given two vectors of attributes $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$, the cosine similarity, $\cos(\theta)$, is defined as

$$\cos(\theta) = \frac{\boldsymbol{x_1} \cdot \boldsymbol{x_2}}{||\boldsymbol{x_1}||_2 ||\boldsymbol{x_2}||_2}.$$

A similarity of 1 means two vectors have exactly the same orientation, indicating that the two pieces of texts that produce the vectors have the maximal degree of similarity. And a cosine that is close to 0 indicates that two vectors are nearly perpendicular to each other, reflecting a low degree of similarity of the texts underlying the two vectors.

To perform cosine match, for each unit in the treated group we calculate the cosine between its summary vector and the summary vector of every unit in the control group. After obtaining all the cosines (in total 238×655), in principle we can start our matching in any ordering of the treated units. However, to avoid scenarios where some treated units are matched to control units that share low similarities due to the fact that highly similar control units have already been matched with previous treated units, we ordered our treated units in the following way before matching. For each treated unit, we identify the top 9 control units (with replacement) that are most similar to the treated unit, and calculate the difference between the largest cosine and the smallest cosine among the 9 cosine values, denoting by *cosine difference*. We then sort our treated units by descending order of their cosine difference.

Hence, the matching starts with the treated unit that has the biggest cosine difference among its top 9 cosine values. The rationale behind this is that if the cosine difference is small, the effect of matching to the 9th best choice will not differ much from the effect of matching to the top choice for this treatment unit. However, if the cosine difference is large, then we want this treated unit to be matched with its top choices, as low rank choices will differ a lot from top choices. After applying matching in this order, we get all the matched pairs. Among matched pairs, cosines range from 0.883141 to 0.971589. Such high values of cosine for all the pairs indicates that the matching is potentially good.

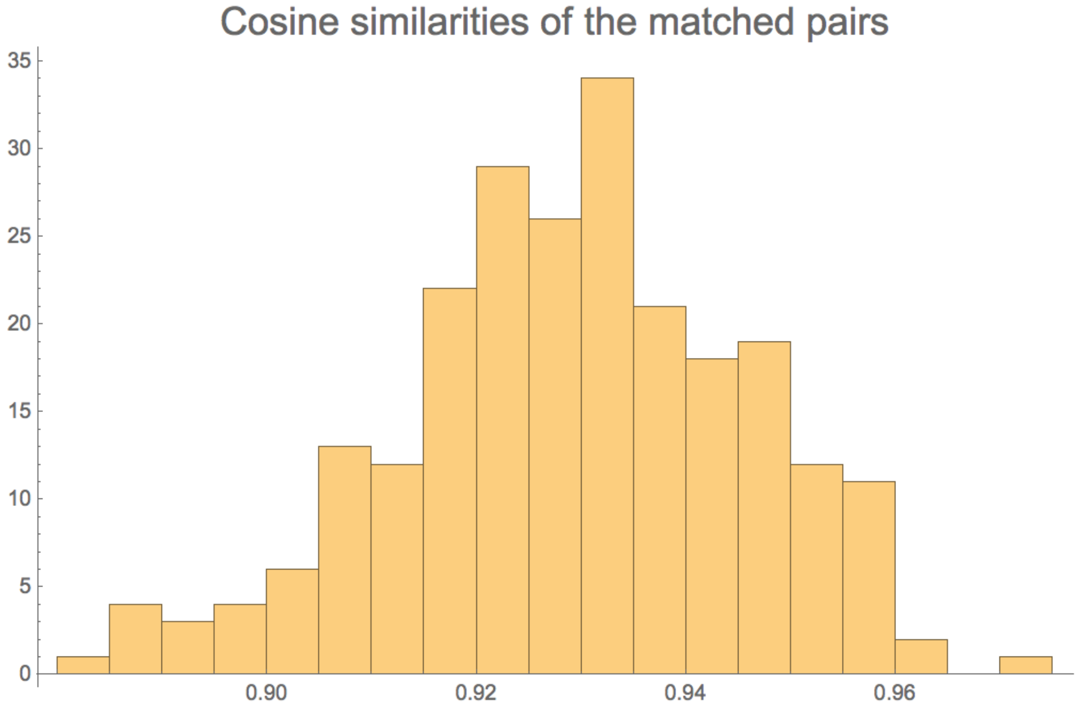

Figure 4: Histogram of cosine values of the matched pairs.

We did not employ two popular distance measures to perform our matching: Mahalanobis metric matching, and propensity scores matching. It has been noted that the Mahalanobis distance works well under situations where the number of covariates is small, but it does not perform as well when there are many covariates. This issue has been discussed in *Stuart 2010* [4] , which is one of the class materials. In our case, these summary vectors are of dimension 300. Hence matching based on Mahalanobis distance is not a good choice. As for propensity score matching, an accurate estimate of propensity score is essential. To estimate propensity score, we need to identify the set of covariate and interactions to use. However due to the availability of data, we only have very limited number of covariates, which may lead very inaccurate estimate of propensity score. And also for the covariate, summary vector, its high-dimensional nature also poses some difficulty in performing the logistic regression well.

After obtaining matched samples, we check covariate balance between the treatment group and selected units from the control group for the rest of the covariates.

If we detect imbalance in any covariate distributions, we will consider employing a two-stage matching method: first, we will perform a one-to-multiple matching using cosine distance, and second, we will try to reduce the already matched samples to a one-to-one pair using Mahalanobis metric or propensity score, which are estimated using only the rest of the covariates available.

The covariates we check are number of authors, and number of months passed since the initial publication of the paper. Number of authors may affect citation counts in various ways. For example, with more authors, a paper will got publicized more frequently, thus drawn more attention, which may potentially be positively related to citation counts. How long a paper has been out is also an important covariate, as we expect citation counts to be non-decreasing in time. We compare the covariate distributions by looking at summary statistics like mean and variance, and also inspect the histogram plot and boxplot of the distributions.

5 Results and Discussion

Using cosine matching we are able to match each treated unit to a unique control unit; and among all matched pairs, the cosine similarities range from 0.918532 to 0.979028. As all cosine similarities are very close to 1, this indicates among a matched pair, their publications share lots of similarities in terms of content, structure, semantics, length etc.

Next we check balance in covariate distributions. Firstly, we compare number of authors. In the treatment group, the mean number of authors of a publication is 5.19, and the variance is 30.27. The control group has mean 5.08 and variance 30.79. Hence the central and dispersion of these two groups are quite similar. We also calculated mean and variance for the entire control group (before matching), and we can see the balance is improved. We also plot overlapping histograms to inspect the overall similarity of the two distribution, as well as boxplots to compare the quantiles. The histograms show that the two distributions are quite balanced at all values, and boxplots shows that the overall dispersion are almost the same, and the three common quantiles are also quite similar.

	Female	Matched Male	Male
Mean	5.19328	4.92017	4.3145
Variance	30.7895	34.5379	21.2832

Table 1: Means and variances of the number of authors.

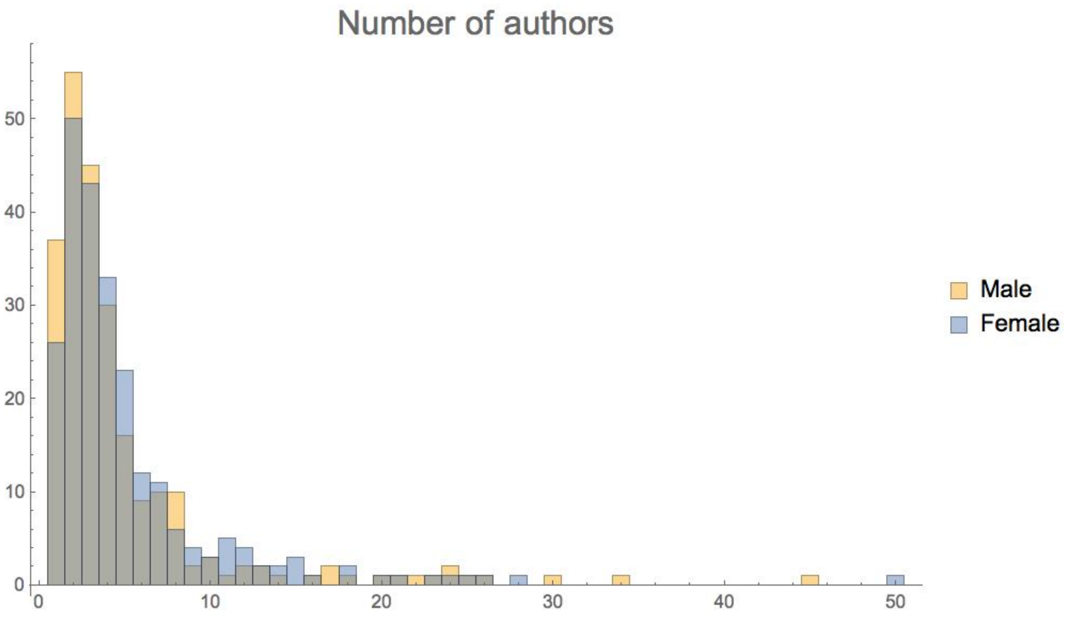

Figure 5: Histogram of number of authors

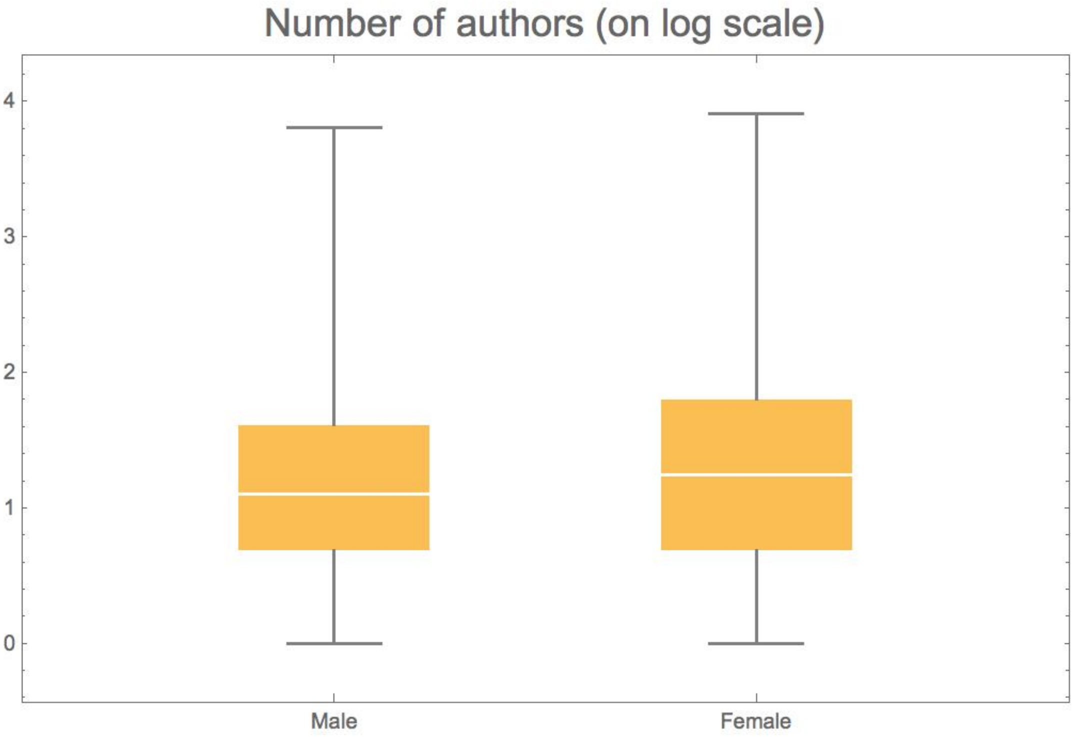

Figure 6: Box-plots of number of authors

The next covariate we check is number of months past since initial publication of a paper. The mean number of months since publication in the treated group is 76, i.e. 6 years and 4 months, with a variance of 98. For the control group, the mean is 77 and variance is 106. This covariate is balanced before matching, as we would expect because they are from the same time period. The matching maintains the balance. Histogram and boxplot comparisons also do not suggest severe imbalance in two distributions.

	Female	Matched Male	Male
Mean	76.042	77.3487	76.6183
Variance	98.1754	105.84	102.512

Table 2: Means and variances of the number of months since publication.

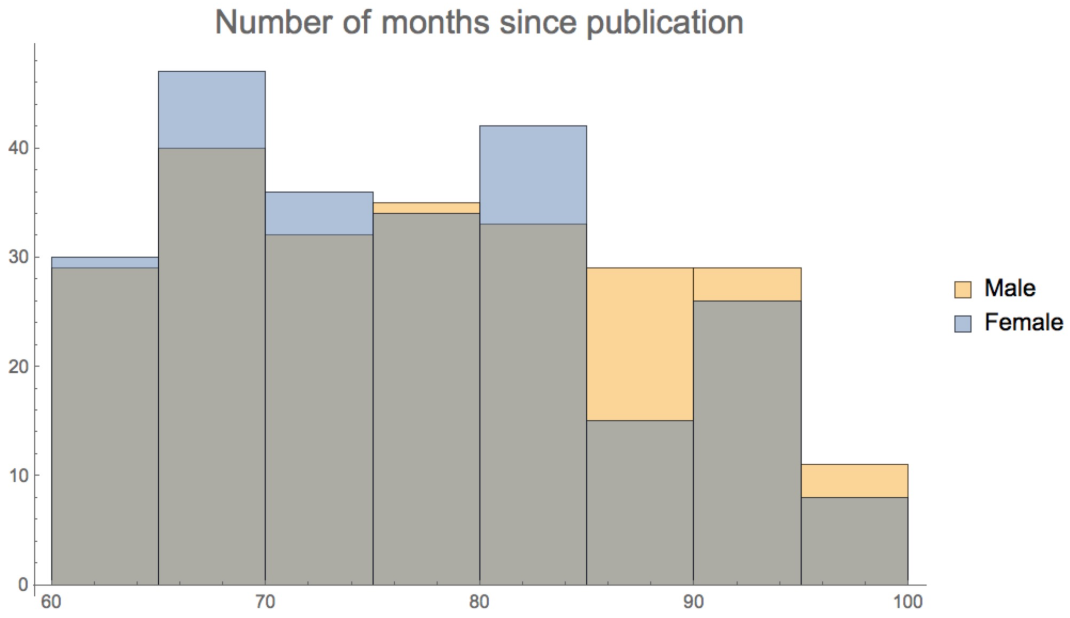

Figure 7: Histogram of number of months since publication in both groups

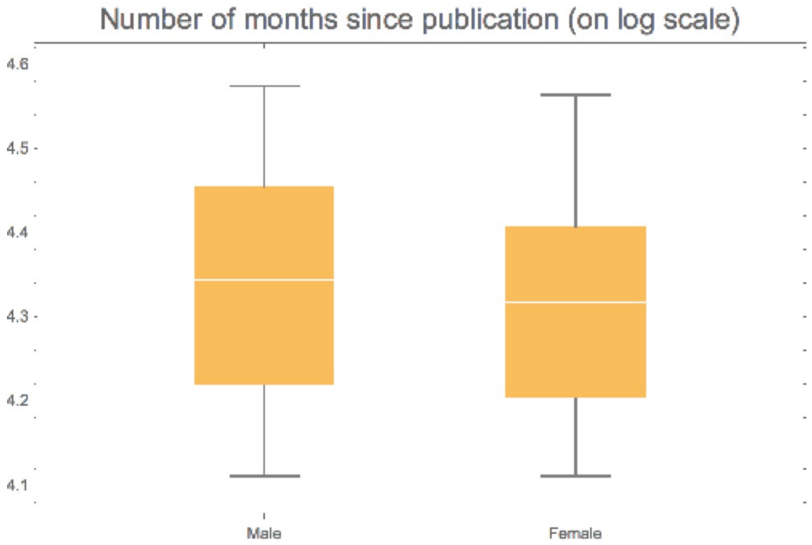

Figure 8: Box-plots of number of months (log-scaled) since publication in both groups

Since the treatment group and control group are quite balanced in all covariates' distributions. We proceed to estimate the average treatment effect on the treated group. The ATT is estimated to be -7.39076. This implies that a paper first-authored by a female is receiving about 7 fewer citations on average compared to otherwise similar paper but published by a male first author. Digging deeper, the average citation counts of those matched male-first-authored papers is 39.5714, compared to the average citation counts of all male-first-authored papers being 35.8489. This means that the male-first-authored papers matched with female-first-authored papers are the ones getting more citations, hence of higher quality. This could potentially imply that a paper with relative lower quality has lower probability of getting published if the first author is a female.

6 Comparing to STM

As structural topic model (STM) is a very popular method to summarize text document, We applied STM to our data as well to compare its perform to that of doc2vec. After some preliminary exploration, we choose to specify 14 topics, and use the "stm" package in R to estimate the model and calculate topic weights for each of the documents we have at hand. Below is a summary of the topics, and some of its popular words.

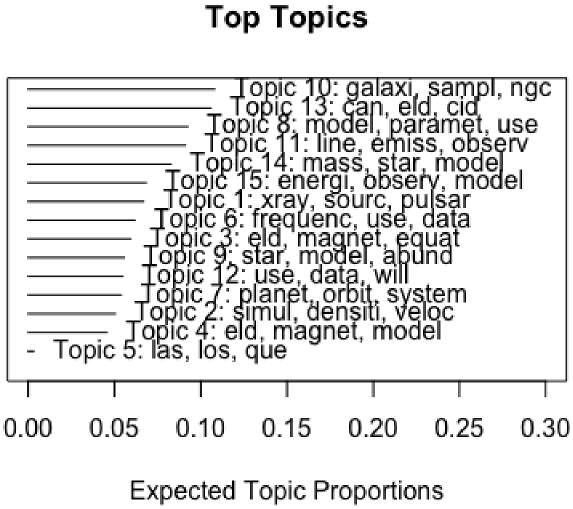

Figure 9: List of Top Topics

Once we have the 14-dimensional topic weight vectors for each document, we perform matching in the same spirit as section 4: we first did cosine matching based on topic weight vectors, and then check other covariate balances.

We again check the balance for the number of months and number of authors. Means, variances, histograms and box-plots are given below.

	Male	Female
Mean	75.979	76.042
Variance	95.0249	98.1754

Table 3: Means and variances of the number of months since publication in treatment and control groups.

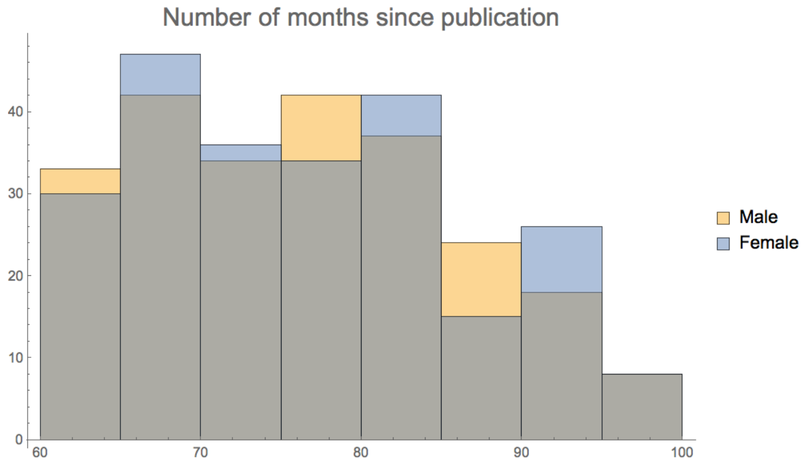

Figure 10: Histograms of numbers of months since publication in both groups.

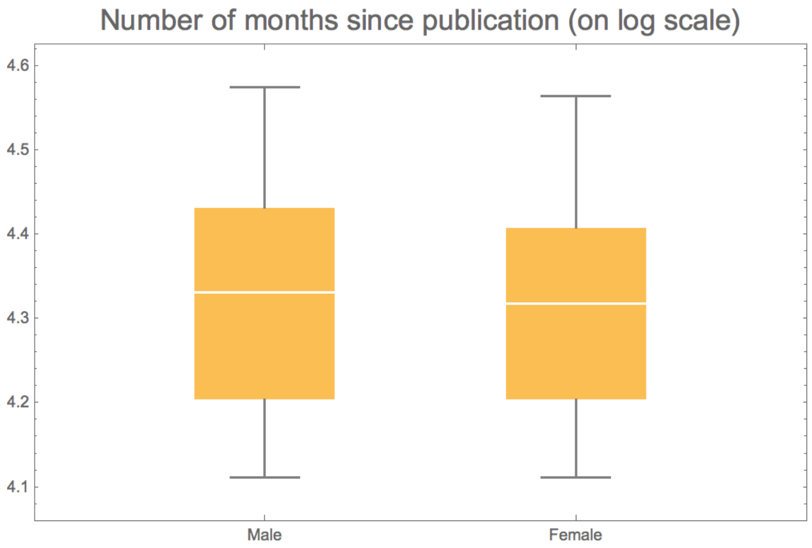

Figure 11: Box-plots of numbers of months since publication (on log scale) in both groups.

	Male	Female
Mean	4.36975	5.19328
Variance	27.5505	30.7895

Table 4: Means and variances of the number of authors in treatment and control groups.

We can see that the covariate: total number of authors, is not well balanced, especially from the perspective of variance. This is due to the outliers in the control group, which is seen more clearly from the histogram and the box-plot. Therefore, doc2vec matching is better than the topic model to some extent.

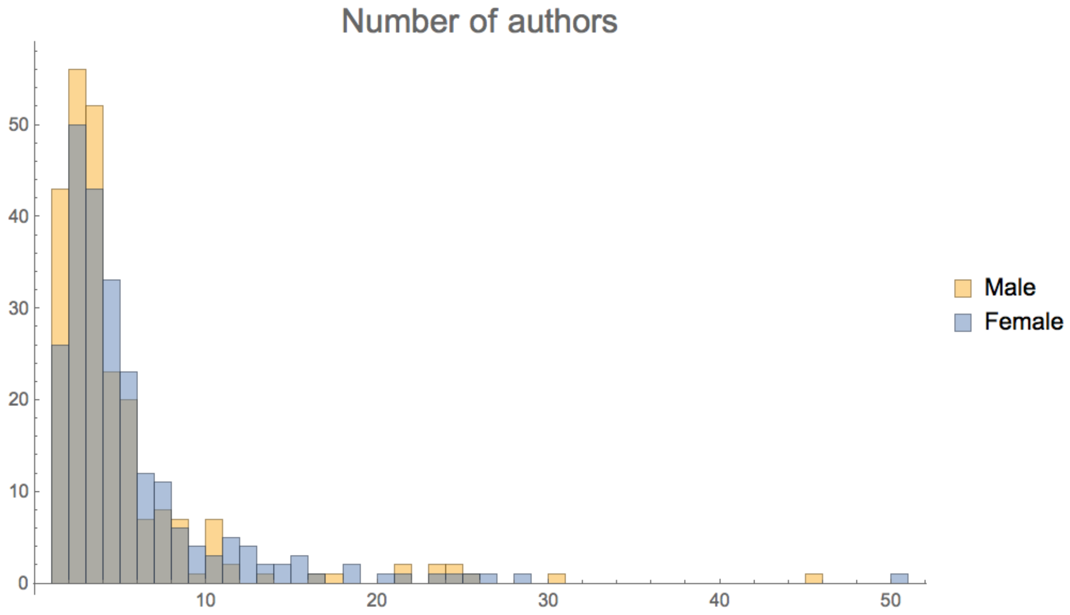

Figure 12: Histograms of numbers of authors in both groups.

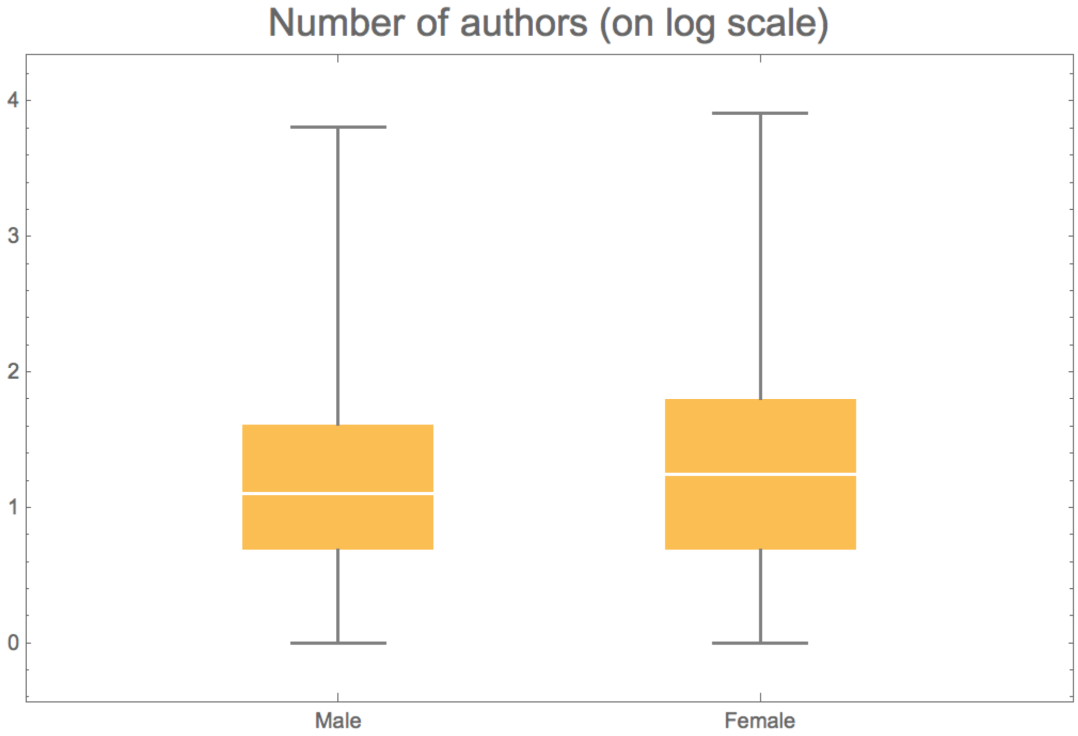

Figure 13: Box-plots of numbers of authors in both groups.

The estimated ATT using topic model matching is -7.09244, similar to the result given by doc2vec. The average citation counts of those matched male-first-authored papers is 39.2731, compared to the average citation counts of all male-first-authored papers being 35.8489. These are in accordance with the results given by doc2vec.

Unfortunately, it is not possible to compare which causal estimates are "correct" and the only possible evaluation of the outcome is to check the covariate balances [4]. That said, given that both matching based on Doc2Vec and STM generate a matching with balanced covariates, both estimates can be said to be valid. Furthermore, the fact the the estimates are quite similar convinces us that there's a gender bias in astronomy publishing.

7 Potential Extensions

1. We were not able to quantitatively claim that matching based on document embeddings is *better* than matching based on other methods e.g. STM. This is very difficult because potential outcomes are unobservable by definition and hence the correct causal effect is unobservable. One way to define what it means for a matching method to perform *better* is to use humans. We can ask Mechanical Turkers to score the similarity of matched documents. This might be somewhat valid if the matched documents are short and not technical because Turkers criteria is likely to coincide with those of astronomy researchers. However, in an application like ours, where the documents are long and highly technical, this is not the case. Hence, although we qualitatively argued that document embedding method is novel in the sense that it captures latent features that are not only topics, we couldn't claim if that results in more accurate estimate of causal effects.

2. Furthermore, in our application, we are using different covariates for the same matching method, not the same covariates for the different matching methods. This is a fundamentally different problem which hasn't been explored in matching literatures since the covariates had been structured in the past. The ways to compare different covariates generated from unstructured data is still left to be researched.

3. To obtain the full text of the papers, we relied on an open source tools that allows us to parse PDFs of those papers. However, because formula and graphs appear often in some of those PDFs, they were converted to gibberish. Perhaps we could've obtained the latex sources of each paper had they been available, but there's broader field of matching on formula and graphs, that are not fully developed in contrast with simple text. Even for document embeddings, it is unclear if latex sources can/should be treated just like pure texts.

8 Conclusion

In this project, we investigated the difference in citation counts of academic publications caused by the perceived gender of the first author in the area of astronomy. In order to create good match, we took the full text documents of the publications into consideration. We took an innovative step to employ the document embedding method, doc2vec, to reduce the dimension of full articles. We also explore the alternative of using STM to summarize full documents. After matching, our two groups are quite balanced in all the covariates we have. Our analysis has shown that publications with "female" first author on average get 7 fewer citations than that of a "male" first-authored paper. This is consistent with existing literature. In the future, we will consider collecting more data on other relevant covariates, e.g. whether the first author has tenured or not, what's the H-index of the first author at the time of publications etc. We hope that with more covariates added in, we will arrived at more accurate estimate. We would also like to explore other methods in high dimensional text matching.

References

[1] Moss-Racusin, C., Dovidio, J., Brescoll, V., Graham, M., Handelsman, J., 2012,. *Science faculty's subtle gender biases favor male students.* Proceedings of the National Academy of Sciences, USA, 109, 16474.

[2] Maliniak, D., Powers, R., Walter, B., 2013. *The Gender Citation Gap in International Relations.* International Organization 67, 2013.

[3] Caplar,. N., Tacchella, S., Birrer, S., 2016 *Quantitative evaluation of gender bias in astronomical publications from citation counts* Nature Astronomy, 2017

[4] Stuart, Elizabeth 2010 *Matching Methods for Causal Inference: A Review and a Look Forward* Statistical Science, 2010

[5] Quoc Le, Tomas Mikolov 2014 *Distributed Representations of Sentences and Documents* International Conference on Machine Learning, 2014

[6] Jeffrey Pennington, Richard Socher, Christopher D. Manning 2014 *GloVe: Global Vectors for Word Representation*

[7] Chris McCormick 2016 *Word2Vec Tutorial - The Skip-Gram Model* http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

[8] Margaret E. Roberts, Brandon M. Stewart, and Richard Nielsen 2016 *Matching Methods for High-Dimensional Data with Applications to Text*