Vectors of Teaching and Research in Texan Universities and Cities

Bhargav Srinivasa Desikan[†]

[†]Knowledge Lab, University of Chicago

Introduction

What is the complex relationship between **research output** and **teaching intent** among universities and cities across the USA? *How* is the research conducted in an university related to the research conducted in other universities? *How* is the teaching conducted in an university related to the teaching in other universities? And *how* are research and teaching related - within universities and between universities?

To begin to be able to answer these questions, a merely qualitative response is not only inadequate, but impossible to verify. A thorough textual analysis is required, and one which uses massive data sources to be able to have any kind of generalizability. In my research I will attempt to curate the data which will allow us to begin to answer these questions, as well as lay the methods down which will give us the ability to navigate this space. I will also present preliminary results and future directions.

Data

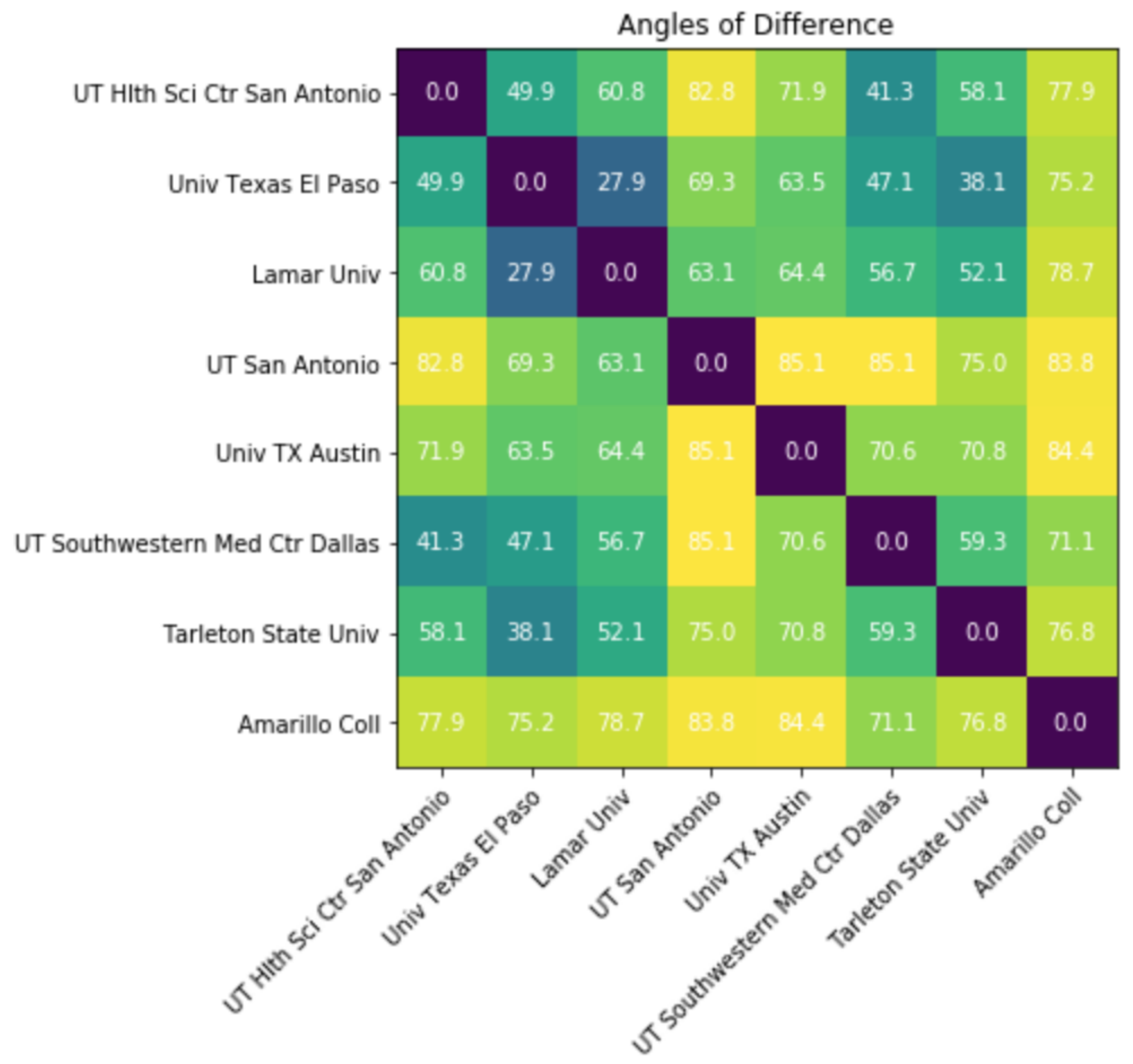
We are in an unique position to be able to answer these kind of questions because of two very rich data sources. One is the **Web of Science** dataset, which contains metadata of research papers from 1900-2016, including organisation, authors, and importantly for the purpose of my research question, paper abstracts. These paper abstracts help us to quantitatively measure the qualitative nature of the paper. The dataset is very rich, but we will only be using four facets of this information: the **abstract** of the paper published, the **cities** associated with the paper (a paper can be associated with multiple cities), the **organisations** the paper is associated with (similarly, multiple organisations), and the **year** the paper was published in.

The second data set is from the **Open Syllabi Project**, which contains syllabi data. As the data is largely mined off the internet and from multiple sources, it is unstructured and required thorough preprocessing before being used for any analysis. Again, we will use only four features of each of the syllabi from this dataset - the **textual/numeric contents of the syllabus**, the **city** associated with the syllabus, the **institution** associated with the syllabus, and the **year** the syllabus is intended to be used. While research papers normally have multiple associations, syllabi tend to not.

These datasets will allow us to model the research and the teaching being conducted in cities and organisations across the USA. For the purpose of my pilot study, I have collected data only for the state of Texas.

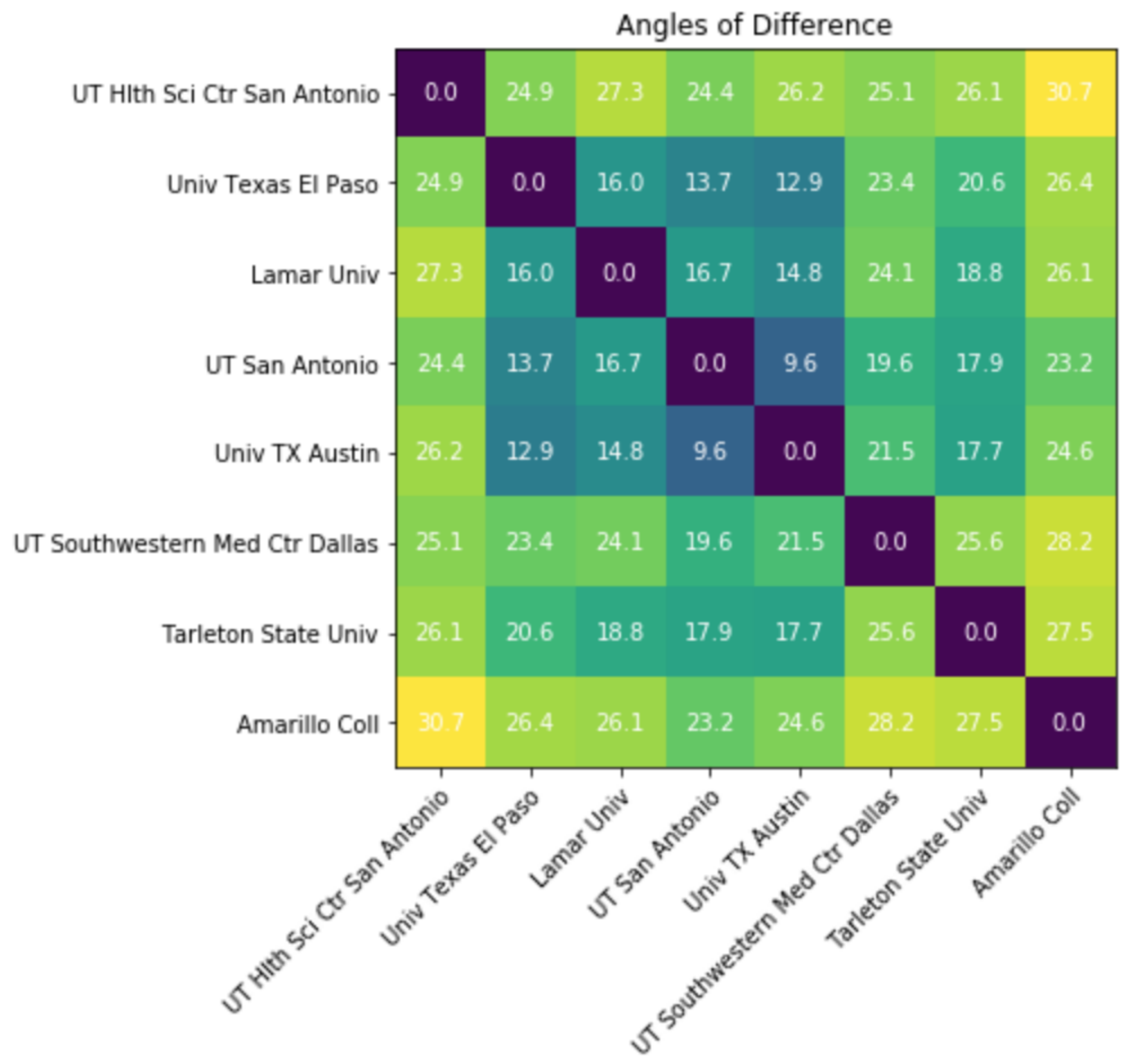
Results

After curation and organisation of the data, we are able to create our city and institution vectors. Our initial results involve analyses of the relationships of research and teaching between and within universities.



	UT Hlth Sci Ctr San Antonio	Univ Texas El Paso	Lamar Univ	UT San Antonio	Univ TX Austin	UT Southwestern Med Ctr Dallas	Tarleton State Univ	Amarillo Coll
UT Hlth Sci Ctr San Antonio	0.0	49.9	60.8	82.8	71.9	41.3	58.1	77.9
Univ Texas El Paso	49.9	0.0	27.9	69.3	63.5	47.1	38.1	75.2
Lamar Univ	60.8	27.9	0.0	63.1	64.4	56.7	52.1	78.7
UT San Antonio	82.8	69.3	63.1	0.0	85.1	85.1	75.0	83.8
Univ TX Austin	71.9	63.5	64.4	85.1	0.0	70.6	70.8	84.4
UT Southwestern Med Ctr Dallas	41.3	47.1	56.7	85.1	70.6	0.0	59.3	71.1
Tarleton State Univ	58.1	38.1	52.1	75.0	70.8	59.3	0.0	76.8
Amarillo Coll	77.9	75.2	78.7	83.8	84.4	71.1	76.8	0.0

Fig. 1: Research-Research Comparisons



	UT Hlth Sci Ctr San Antonio	Univ Texas El Paso	Lamar Univ	UT San Antonio	Univ TX Austin	UT Southwestern Med Ctr Dallas	Tarleton State Univ	Amarillo Coll
UT Hlth Sci Ctr San Antonio	0.0	24.9	27.3	24.4	26.2	25.1	26.1	30.7
Univ Texas El Paso	24.9	0.0	16.0	13.7	12.9	23.4	20.6	26.4
Lamar Univ	27.3	16.0	0.0	16.7	14.8	24.1	18.8	26.1
UT San Antonio	24.4	13.7	16.7	0.0	9.6	19.6	17.9	23.2
Univ TX Austin	26.2	12.9	14.8	9.6	0.0	21.5	17.7	24.6
UT Southwestern Med Ctr Dallas	25.1	23.4	24.1	19.6	21.5	0.0	25.6	28.2
Tarleton State Univ	26.1	20.6	18.8	17.9	17.7	25.6	0.0	27.5
Amarillo Coll	30.7	26.4	26.1	23.2	24.6	28.2	27.5	0.0

Fig. 2: Teaching-Teaching Comparisons

We observe that the average angle of difference for research in universities is **61.14°**, for teaching is **26.98°** , between-university research and teaching is **76.94°**, and within-university research and teaching is **77.63°**. In other words, teaching across universities is more similar than research across universities.

Methods

To be able to effectively use this collection of organised textual data, it is necessary to quantitatively represent this. To measure an aggregate **similarity**, **evolution**, or **orientation**, a method such as bag of words or TF-IDF might not be adequate.

For the purposes of this research study I have used Doc2Vec, an extension of the popular Word2Vec **word embedding** method. Word embedding allow us to use a semantically relevant n-dimensional (for our purposes, n=100) vector representation of documents.

This vector representation is *crucial* for our purposes, as it allows us to aggregate all the document vectors for a city, and for an institution. By simply finding the mean of all the vectors of the research paper abstracts associated with a city or organisation, it is possible to get a fairly accurate representation of the kind of research carried out there - the **research vector** of that entity. Similarly, we can calculate the **teaching vector** of the entity by calculating the mean of all the vectors of the syllabi associated with that entity. These research and teaching vectors form the basis of our analysis. My pilot study analysis largely uses similarity between vectors to make inferences about the nature of research and teaching. In particular, I use **cosine similarity** to identify an angle between the vectors of research and teaching for the various cities and organisations in the state of Texas.

Ongoing and Future Work

The results have given us a pattern for the relationship between research & research, teaching & teaching, and research & teaching for entities in the state of Texas - the obvious next step is to conduct this analysis across all the entities in the USA, before moving on to the world. This analysis is currently in the works, with the results being looked at presently.

What about the non-obvious next step?

Identifying the similarities and differences between different classes of universities: how do **different tier** universities differ in research and teaching? Which tier universities tend to have more similarities within their own research and teaching? How does the **geographical location** of the university effect the nature of the research and teaching? Do universities closer to each other teach or research similar things? How does the research and teaching vector of a university change *over time*? All of these questions and more are currently being set-up for answering once I have research and teaching vectors generated across the USA.