Evolution of Topics and Novelty in Science

Omar Ballester*
Orion Penner

9th Global TechMining Conference
Atlanta
17.10.2019

EPFL

Robust similarity measures from topic modeling: validation and use.

Omar Ballester*
Orion Penner

9th Global TechMining Conference
Atlanta
17.10.2019

EPFL

Overview

Topic modeling (and more generally, latent space) approaches are increasingly used to characterize the content of documents within large-scale science and technology data sets.

However, these approaches are not always robust in neither a statistical sense, nor in terms of a users end goal.

Overview

Topic modeling (and more generally, latent space) approaches are increasingly used to characterize the content of documents within large-scale science and technology data sets.

However, these approaches are not always robust in neither a statistical sense, nor in terms of a users end goal.

We develop and apply a methodology for evaluating the statistical robustness of topic models.

And in doing so, we find that the neural-network based doc2vec produces the best results.

Outline

- Background and Motivation
    - word2vec/doc2vec
- Barriers to application
    - Statistical robustness
    - Descriptive power
    - Reflect reality?
- Our work
    - Quantifying robustness
    - Quantifying descriptive power
    - Towards comparisons with reality
- Wrap up

Outline

Background and Motivation

- word2vec/doc2vec

Barriers to application

- Statistical robustness
- Descriptive power
- Reflect reality?

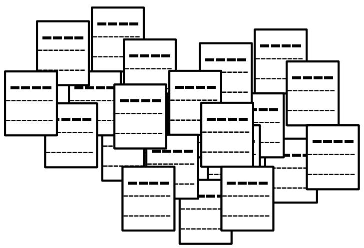
Our work

- Quantifying robustness
- Quantifying descriptive power
- Towards comparisons with reality

Wrap up

Background and Motivation

A **topic model** is a statistical model for extracting the abstract “topics” from a set of documents.



Topic
Model

Science
Patent
Innovation
University
Publication
Firm

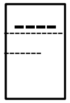
Background and Motivation

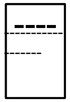
- LDA (Blei, 2003) is a stochastic decomposition based on co-ocurrence probabilities

- NMF (Lee, 1999) is a matrix decomposition using only non-negative factors

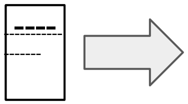
- Embeddings as representations of text (word2vec, fastext, GloVe…)

Background and Motivation

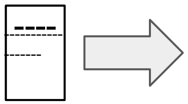




Background and Motivation



Topic
Model



Topic
Model

Background and Motivation

Topic Model

65% - Science
15% - University
30% - Publication

Topic Model

45% - Firm
30% - Patent
25% - Innovation

Robust similarity measures from topic modeling: validation and use.
9th Global TechMining Conference
Atlanta 17.10.2019
O Ballester*, O. Penner

Background and Motivation

Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach
Stephen Hansen, Michael McMahon, Andrea Prat
QJE, 2018

Unsupervised word embeddings capture latent knowledge from materials science literature
Tshitoyan et al.
Nature, 2019

Background and Motivation

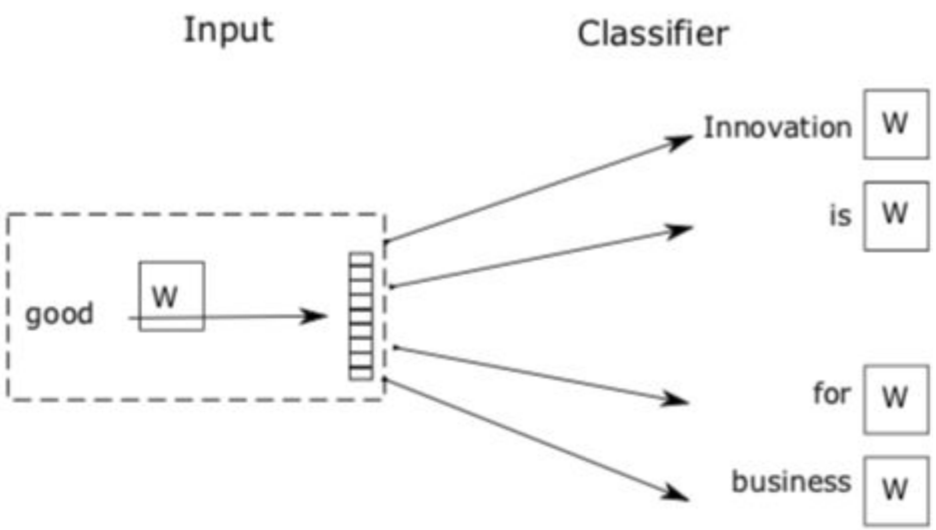
Finding scientific topics
Steyvers, Griffiths
PNAS, 2004

Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches
Boyack et al.
PLOS ONE, 2011

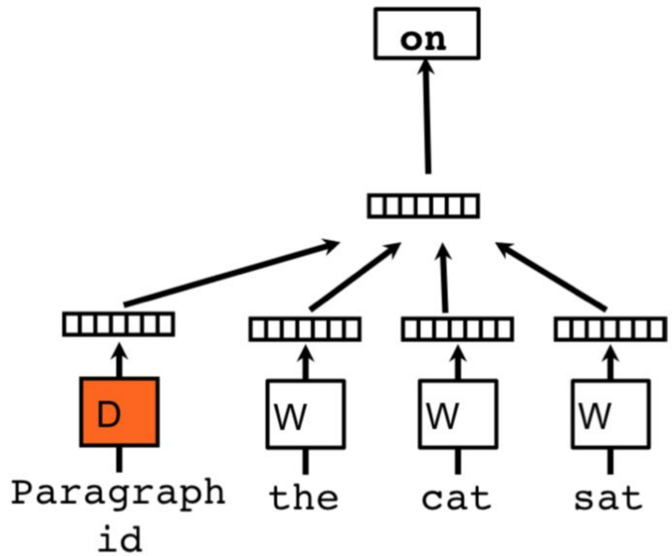
Outline

- Background and Motivation
  - word2vec/doc2vec
- Barriers to application
  - Statistical robustness
  - Descriptive power
  - Reflect reality?
- Our work
  - Quantifying robustness
  - Quantifying descriptive power
  - Towards comparisons with reality
- Wrap up

Word2Vec & Doc2Vec (Mikolov, 2013, 2014)



Skip-Gram architecture. Source: Lenz & Winkler, 2018



DM architecture. Mikolov & Le, 2014

Outline

Background and Motivation

- word2vec/doc2vec

Barriers to application

- Statistical robustness
- Descriptive power
- Reflect reality?

Our work

- Quantifying robustness
- Quantifying descriptive power
- Towards comparisons with reality

Wrap up

Barriers to application

While up-to-date models seem to bring many benefits, they also have some potential drawbacks…

- Bag of words might not be the best representation (Niu et al., 2015)
- Unstable topical representations (Belford, 2017)
- Scalability-interpretability issues (Blei et al, 2007)

Which we try to address…

And other issues: Pre-determined number of topics (Glaser, 2017); Long computation times (Ai, 2016)...

Outline

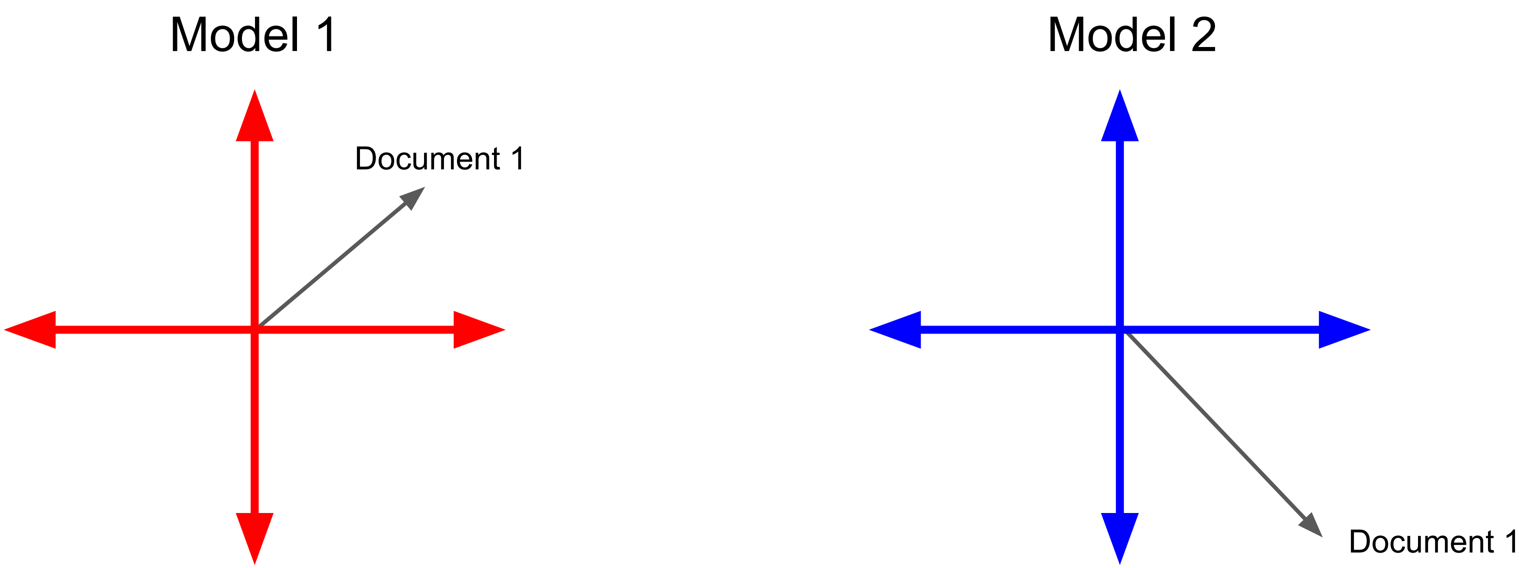
Background and Motivation
- word2vec/doc2vec

Barriers to application
- Statistical robustness
- Descriptive power
- Reflect reality?

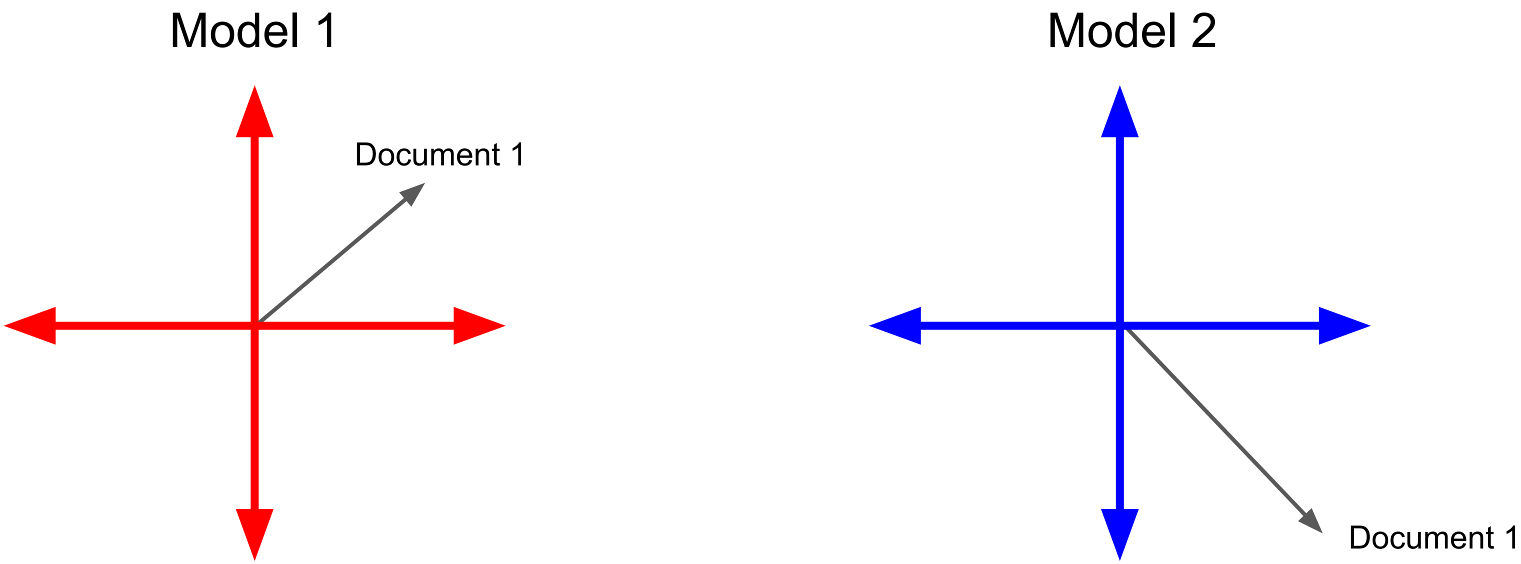
Our work
- Quantifying robustness
- Quantifying descriptive power
- Towards comparisons with reality

Wrap up

Statistical robustness

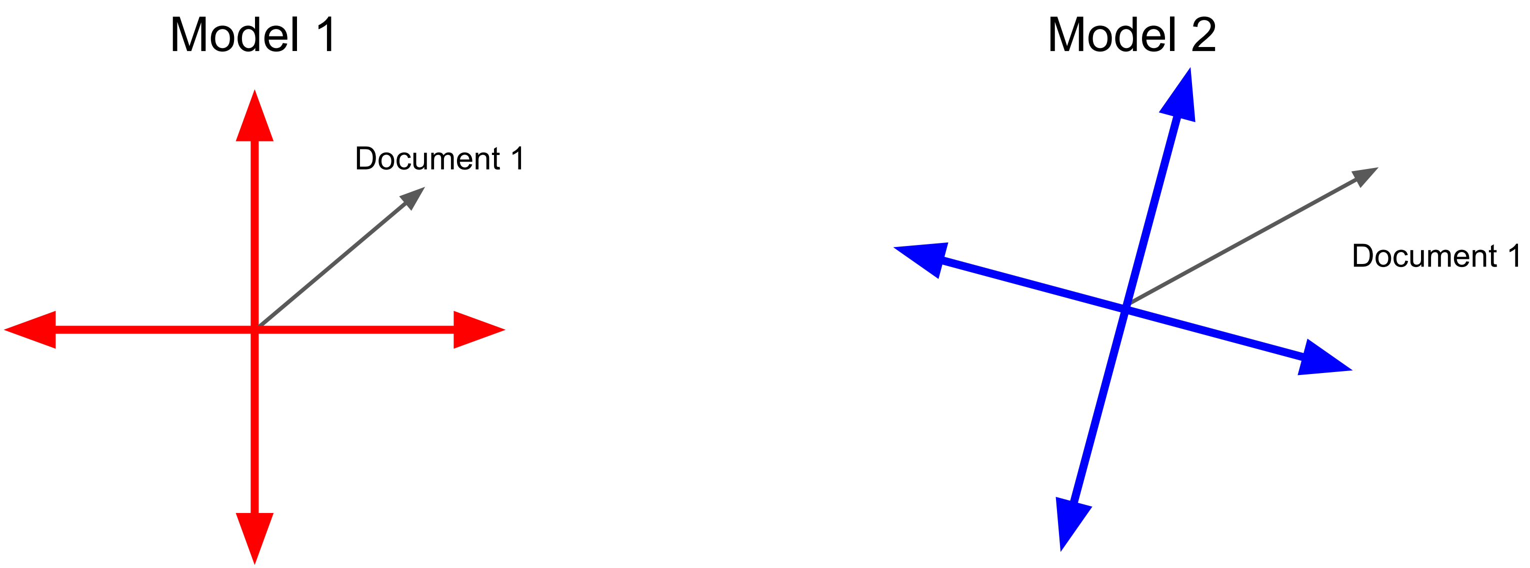


Statistical robustness



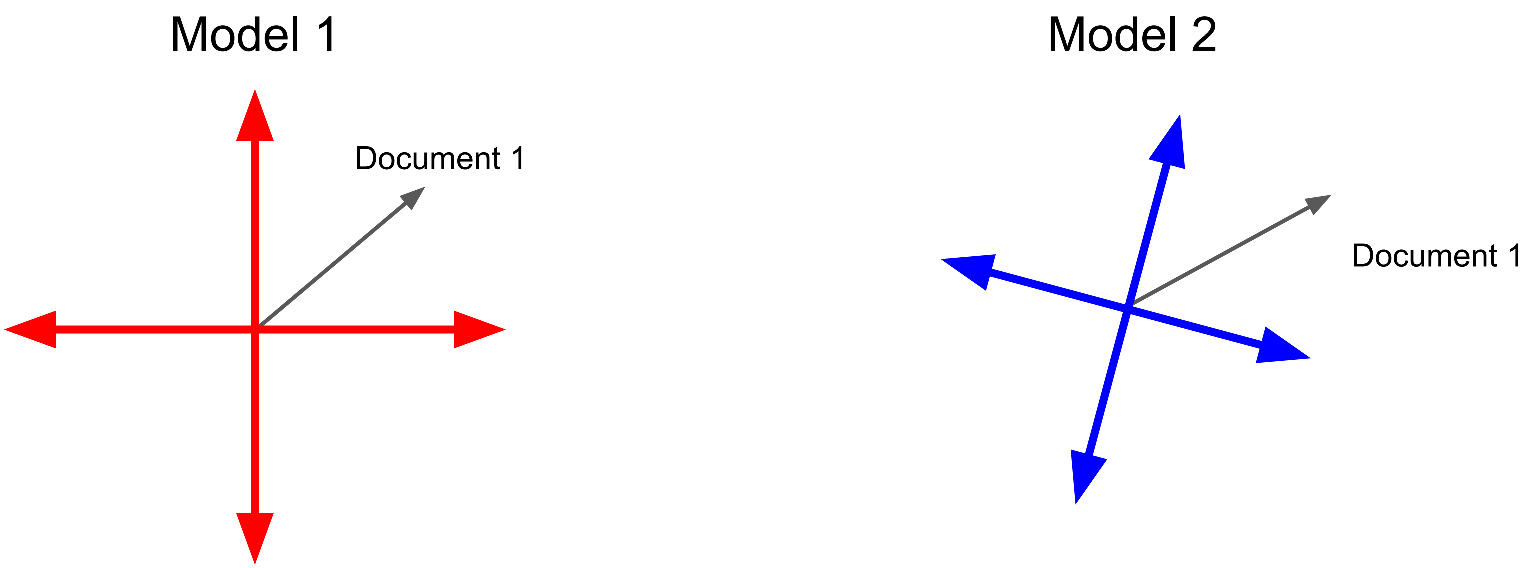
In theory these should agree, but the fact they do not is not, necessarily, a problem. Because the axis are not necessarily the same.

Statistical robustness



In theory these should agree, but the fact they do not is not, necessarily a problem. Because the axis are not necessarily the same. There can be rotations.

Statistical robustness



In theory these should agree, but the fact they do not is not, necessarily a problem. Because the axis are not necessarily the same. There can be rotations. And length scale changes.

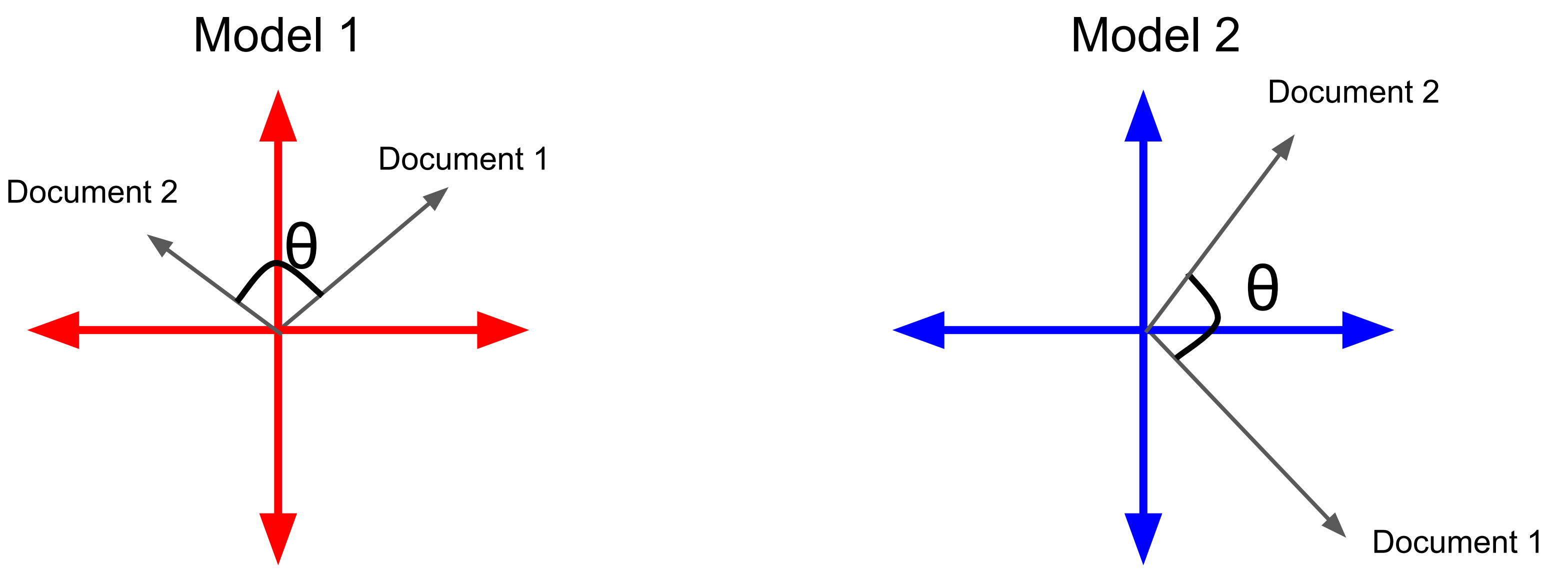
Statistical robustness

So in light of the fact we can have these, essentially arbitrary transformations of the coordinate space, we require a different way of measuring agreement between models.

Statistical robustness

So in light of the fact we can have these, essentially arbitrary transformations of the coordinate space, we require a different way of measuring agreement between models.

We propose cosine similarity.

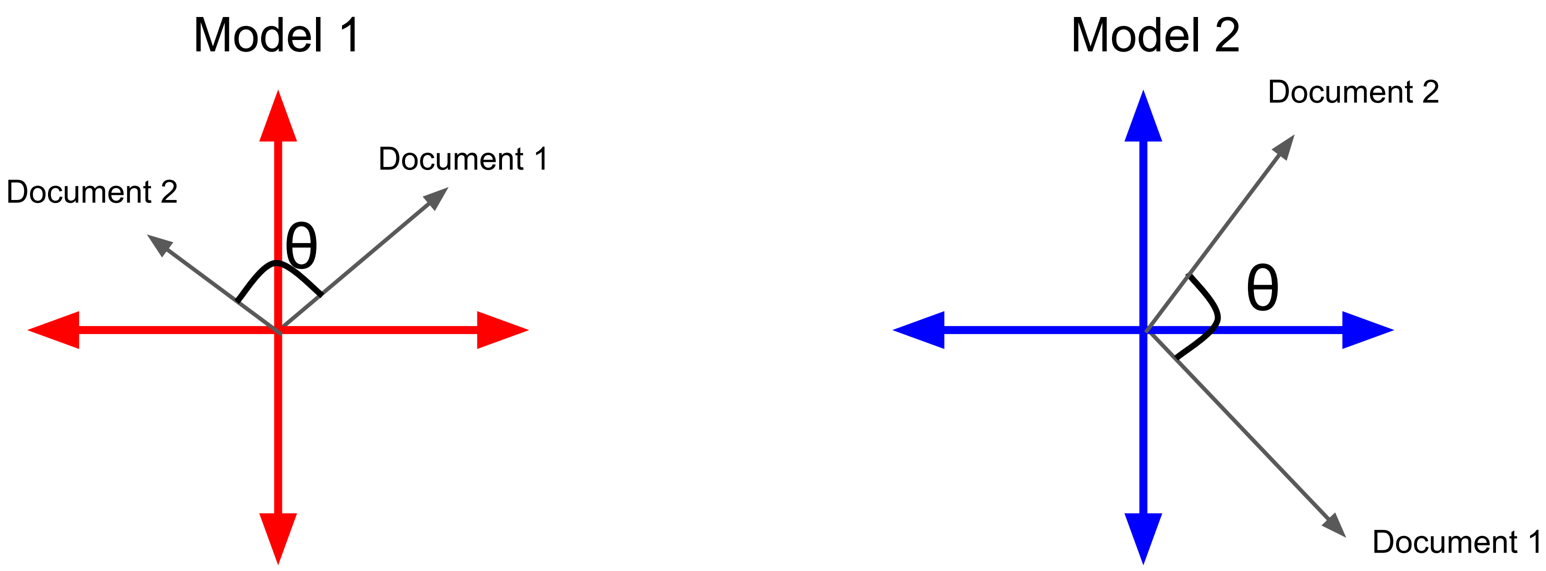


Statistical robustness

The outstanding issue is, however, even when using pairwise cosine similarities most approaches are **not** robust.

Statistical robustness

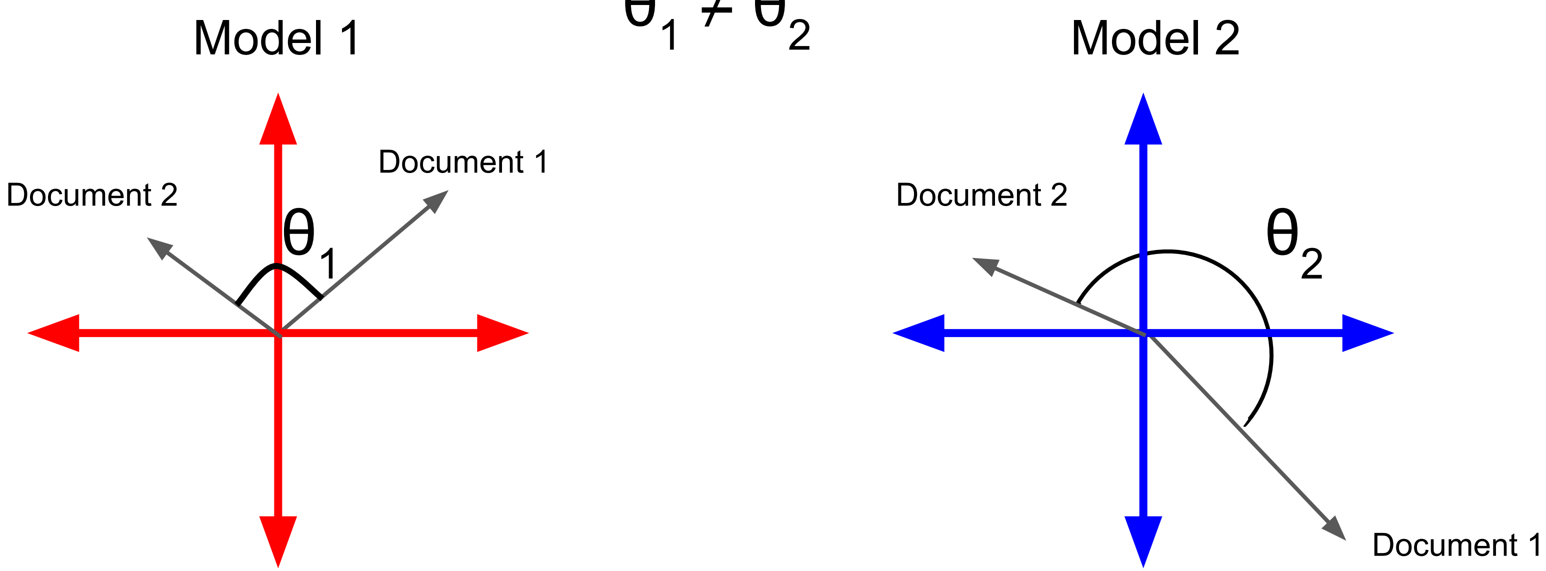
The outstanding issue is, however, even when using pairwise cosine similarities most approaches are **not** robust.



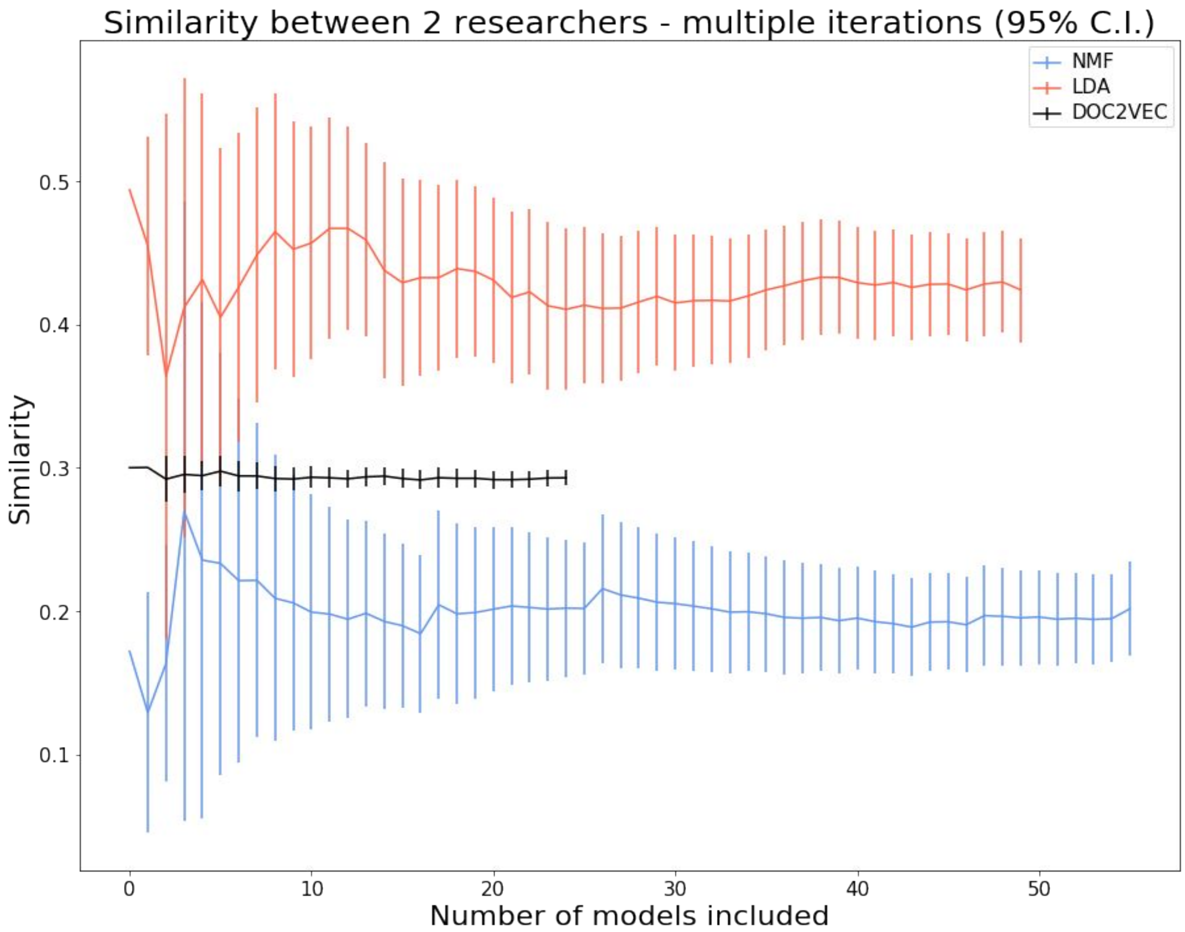
Statistical robustness

The outstanding issue is, however, even when using pairwise cosine similarities most approaches are **not** robust.

$\theta_1 \neq \theta_2$



Statistical Robustness



Outline

Background and Motivation
- word2vec/doc2vec

Barriers to application
- Statistical robustness
- Descriptive power
- Reflect reality?

Our work
- Quantifying robustness
- Quantifying descriptive power
- Towards comparisons with reality

Wrap up

Descriptive power

An additional challenge is that even though one can choose a latent space of any number of dimensions, there is no guarantee that your model is “using” each dimension.

It is largely believed that different latent-sizes serve different purposes (granularity), but that isn’t necessarily the case.

Outline

Background and Motivation
- word2vec/doc2vec

Barriers to application
- Statistical robustness
- Descriptive power
- Reflect reality?

Our work
- Quantifying robustness
- Quantifying descriptive power
- Towards comparisons with reality

Wrap up

Reflect reality

The vectors produced by any approach should ultimately make sense given what we know about science and technology.

Reflect reality

The vectors produced by any approach should ultimately make sense given what we know about science and technology.

- Scientific publications from the same disciplines or fields should be close to each other in the latent space.
- Vectors of patents from the same assignee should be close to each other.
- *etc*.

Reflect reality

The vectors produced by any approach should ultimately make sense given what we know about science and technology.

- Scientific publications from the same disciplines or fields should be close to each other in the latent space.
- Vectors of patents from the same assignee should be close to each other.
- *etc*.

Formal validation is very challenging however as it is relying on some **ground truth** that if we really had to begin with, we probably wouldn't be resorting to these approaches anyways!

Outline

- Background and Motivation
  - word2vec/doc2vec
- Barriers to application
  - Statistical robustness
  - Descriptive power
  - Reflect reality?
- Our work
  - Quantifying robustness
  - Quantifying descriptive power
  - Towards comparisons with reality
- Wrap up

Our work

We try to tackle each of the three barriers covered previously:

1. Statistical robustness
2. Descriptive power
3. Reflect reality?

And on 1. and 2. we have reasonably convincing solutions.

On 3. the work remains open ended.

Outline

Background and Motivation
- word2vec/doc2vec

Barriers to application
- Statistical robustness
- Descriptive power
- Reflect reality?

Our work
- Quantifying robustness
- Quantifying descriptive power
- Towards comparisons with reality

Wrap up

Quantifying robustness

We are looking for an approach that when run multiple times, with different random seeds, produces the same (or highly similar) similarity relationships between the documents.

Quantifying robustness

We are looking for an approach that when run multiple times, with different random seeds, produces the same (or highly similar) similarity relationships between the documents.

Specifically, we are looking for the different runs of the model to produce the **same pairwise cosine similarity**.

Quantifying robustness

We are looking for an approach that when run multiple times, with different random seeds, produces the same (or highly similar) similarity relationships between the documents.

Specifically, we are looking for the different runs of the model to produce the **same pairwise cosine similarity**.

To evaluate this we run a given model and calculate the pairwise cosine similarities between each pair of documents.

Running the model many times with different seeds, we aggregate the cosine similarities (for each pair) and calculate a standard deviation.

Quantifying robustness

We are looking for an approach that when run multiple times, with different random seeds, produces the same (or highly similar) similarity relationships between the documents.

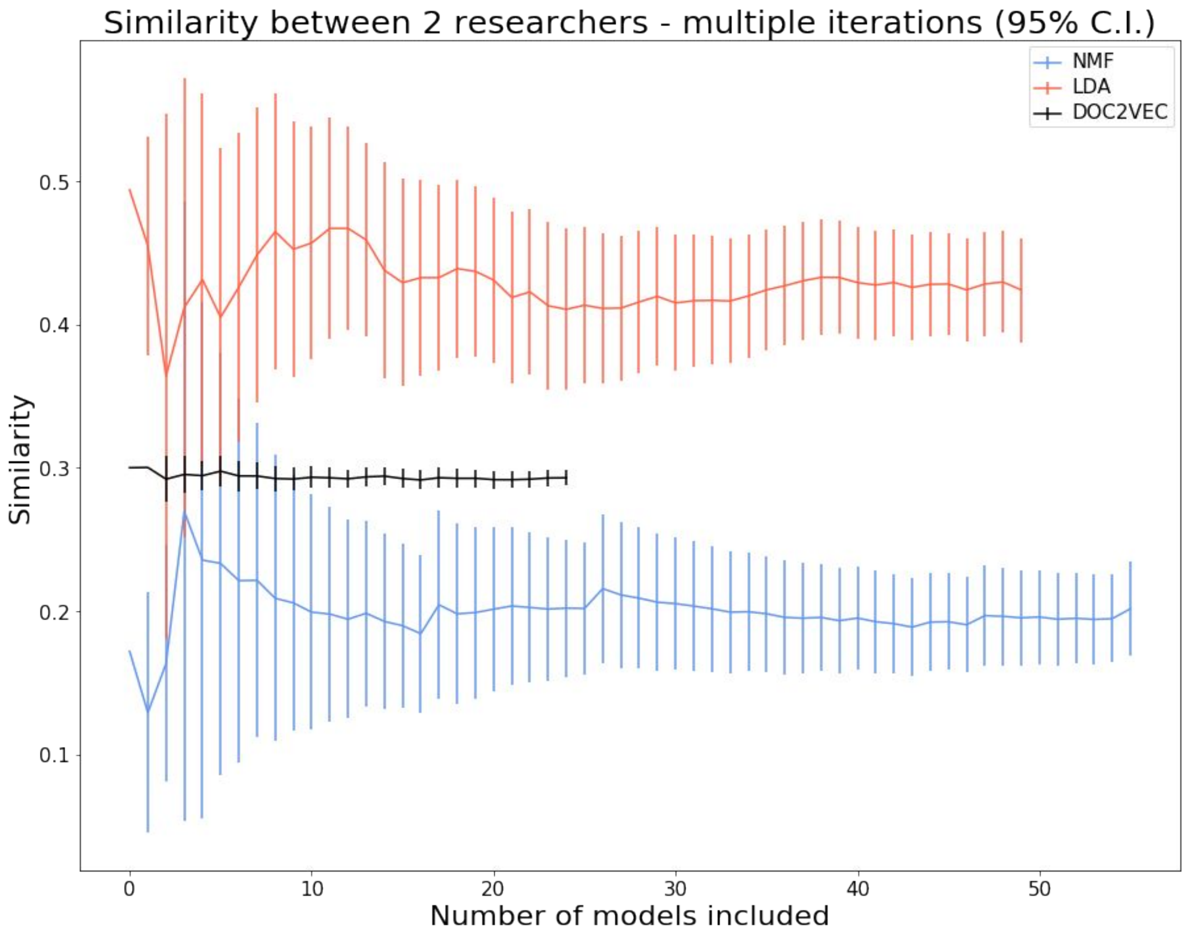
Specifically, we are looking for the different runs of the model to produce the **same pairwise cosine similarity**.

To evaluate this we run a given model and calculate the pairwise cosine similarities between each pair of documents.

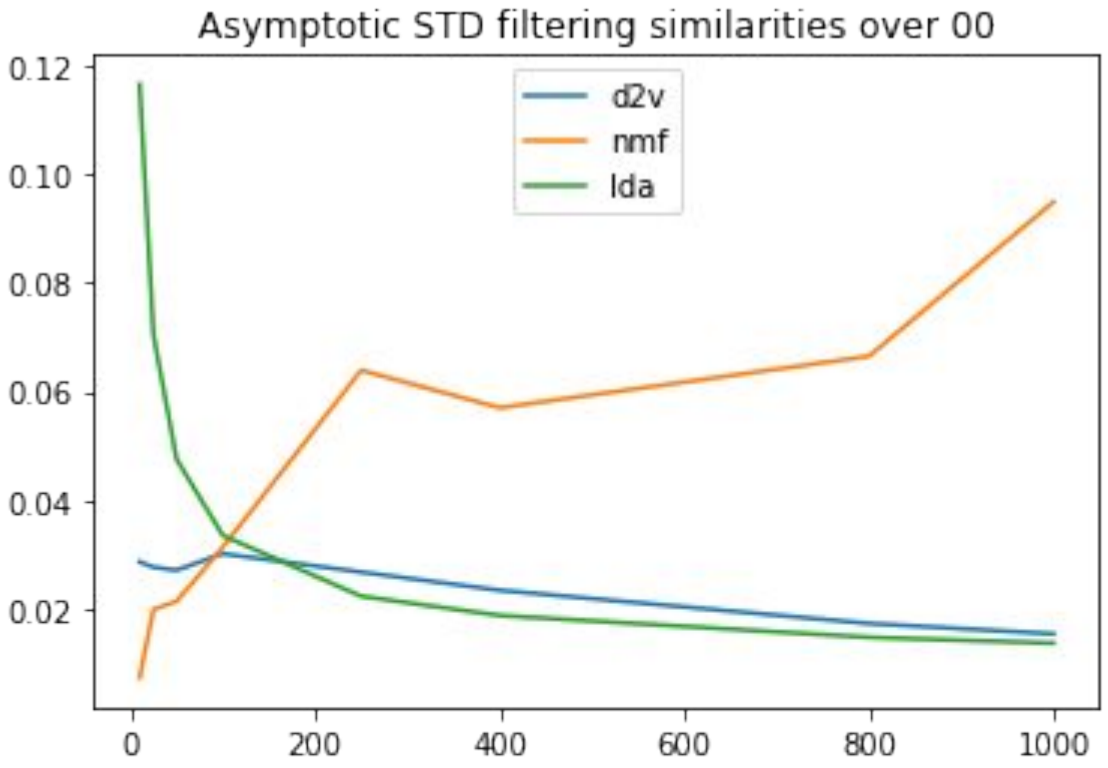
Running the model many times with different seeds, we aggregate the cosine similarities (for each pair) and calculate a standard deviation.

Basically, **how broad is the distribution** of similarities produced for the same pair of documents.

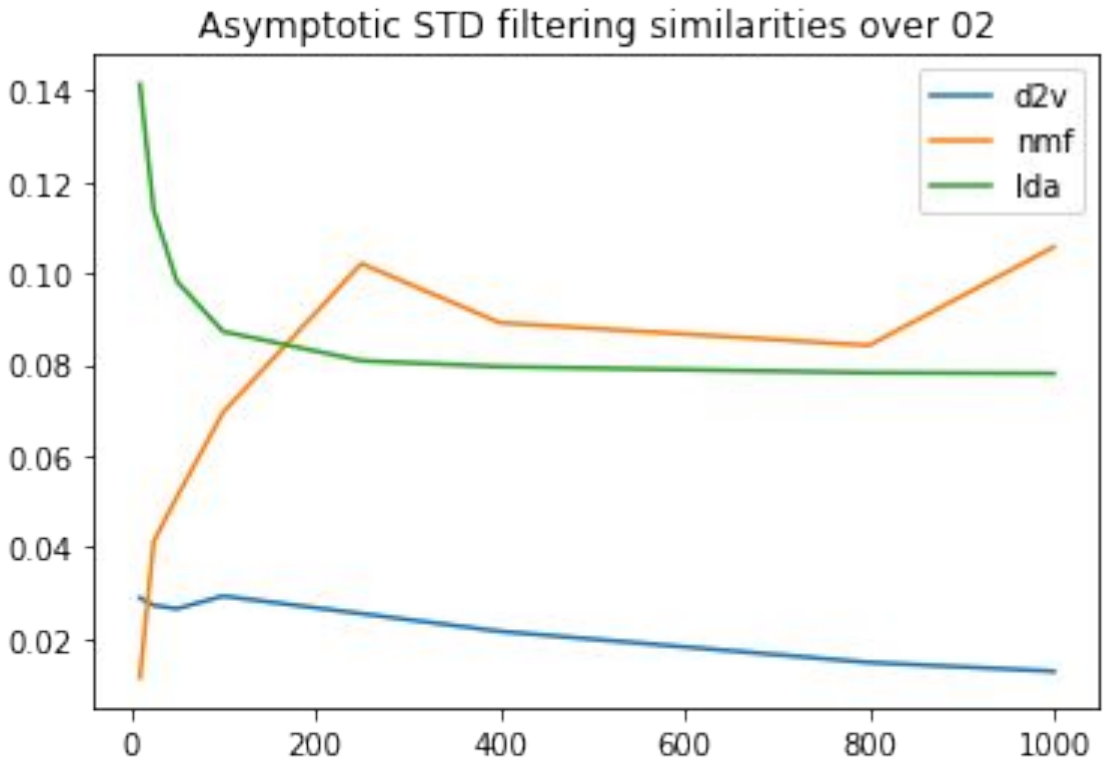
Statistical Robustness



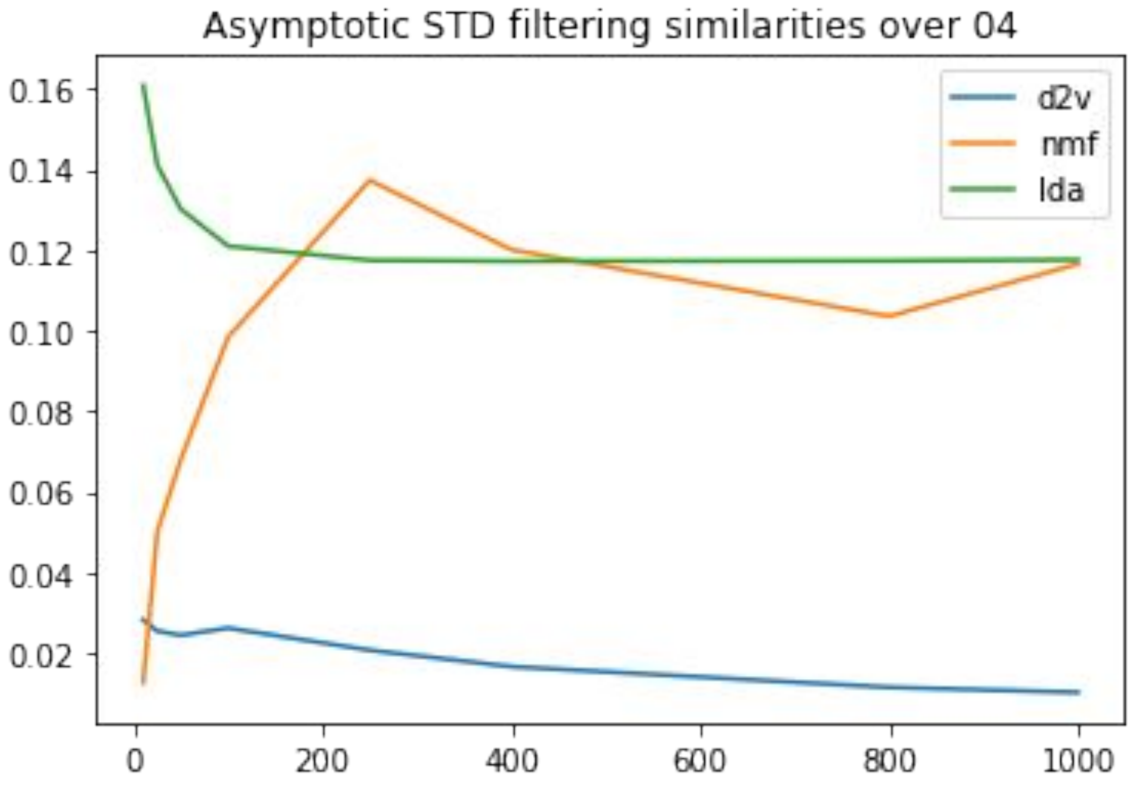
Statistical Robustness



Statistical Robustness



Statistical Robustness



Outline

- Background and Motivation
  - word2vec/doc2vec
- Barriers to application
  - Statistical robustness
  - Descriptive power
  - Reflect reality?
- Our work
  - Quantifying robustness
  - Quantifying descriptive power
  - Towards comparisons with reality
- Wrap up

Quantifying descriptive power

We need to quantify the amount of information that is “gained” on each additional dimension of the model.

Quantifying descriptive power

We need to quantify the amount of information that is “gained” on each additional dimension of the model.

i.e. when we train on 200 topics as opposed to 300 topics, each topic should contain more variation, since **all** the information is reduced to a smaller latent space.

Quantifying descriptive power

We need to quantify the amount of information that is “gained” on each additional dimension of the model.

i.e. when we train on 200 topics as opposed to 300 topics, each topic should contain more variation, since **all** the information is reduced to a smaller latent space.

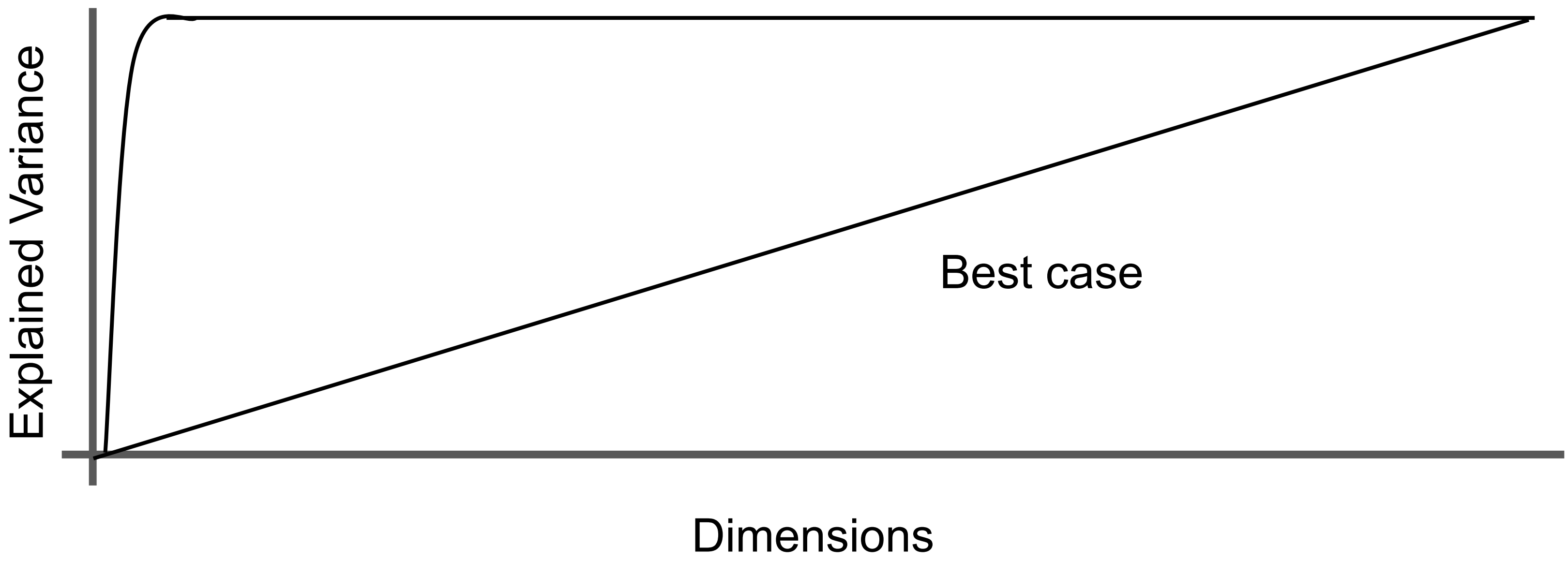
On the other hand, all dimensions of the model should bring value (carry information), so that it makes sense to have a larger latent space for a more fine-grained characterisation.

Quantifying descriptive power

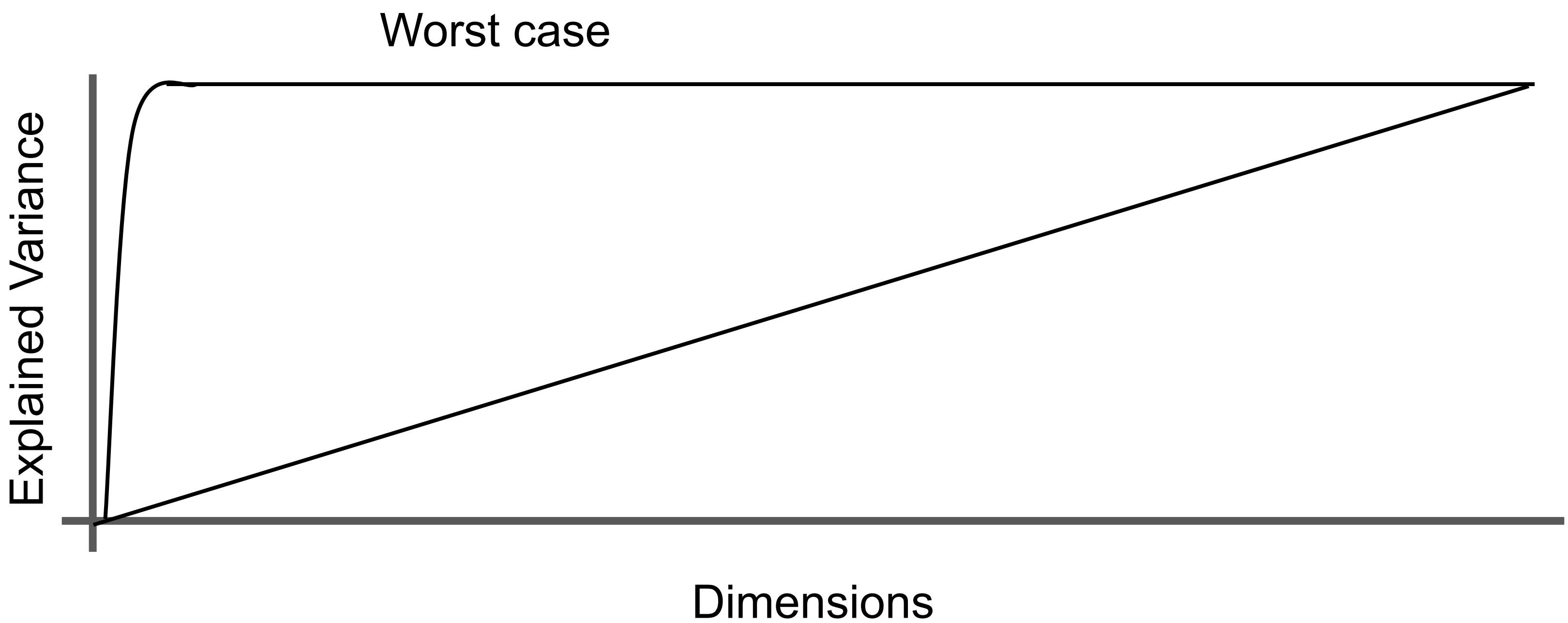
For this, we use **PCA,** a rotation (orthogonal transformation) of the vector-space that yields linearly uncorrelated components **sorted by variance.**

The new rotation has the **same** number of dimensions, ordered from the most informative to the least.

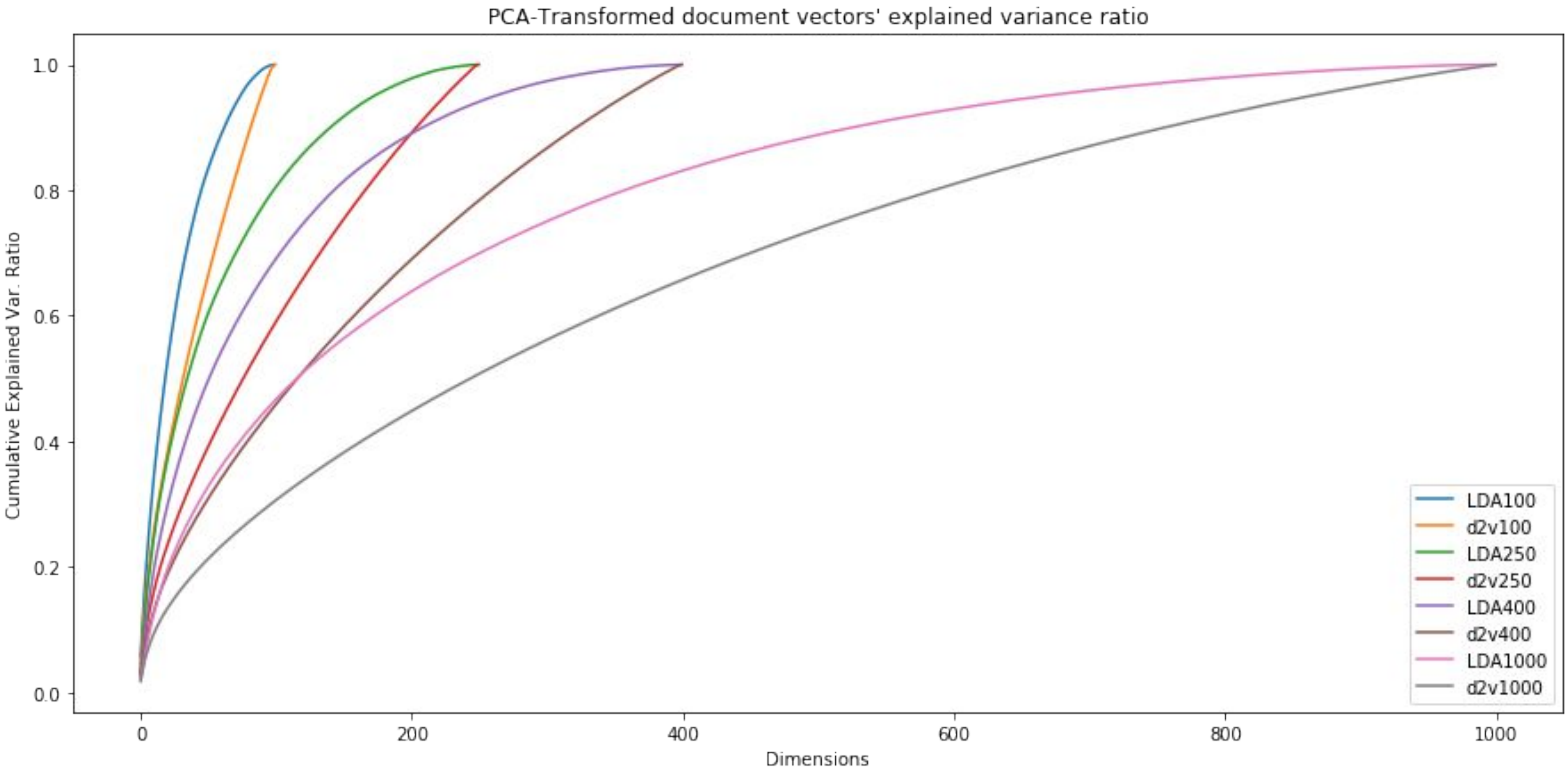
Quantifying descriptive power



Quantifying descriptive power



Scalable Descriptive Power



Outline

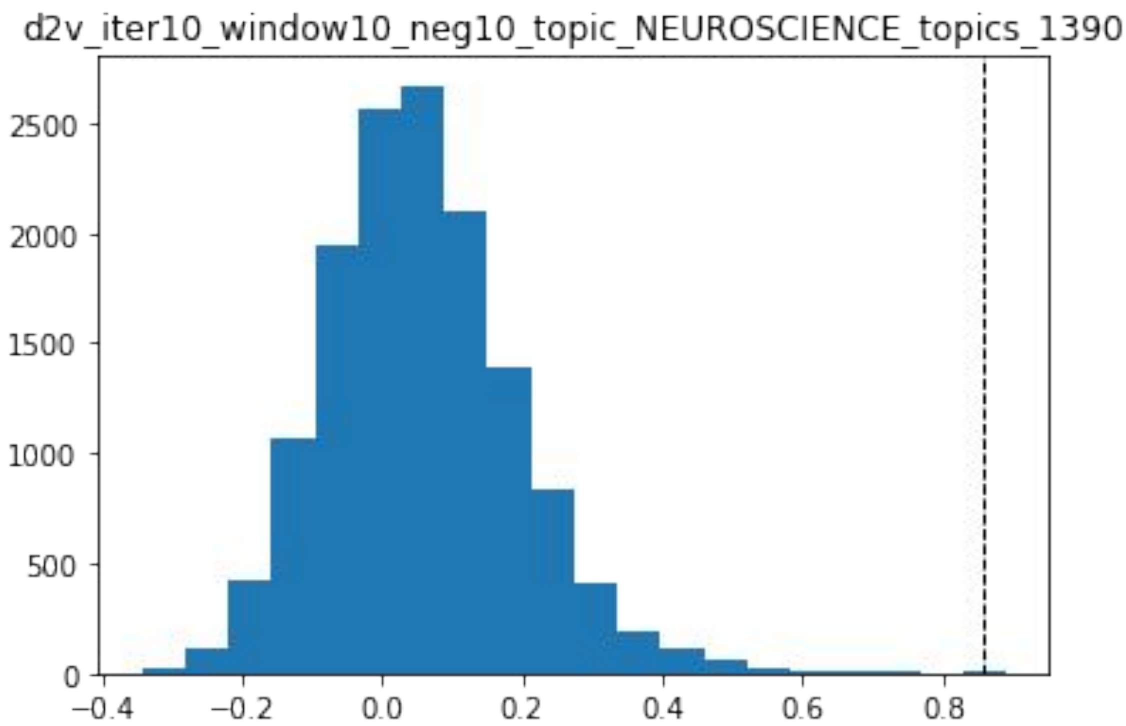
- Background and Motivation
  - word2vec/doc2vec
- Barriers to application
  - Statistical robustness
  - Descriptive power
  - Reflect reality?
- Our work
  - Quantifying robustness
  - Quantifying descriptive power
  - Towards comparisons with reality
- Wrap up

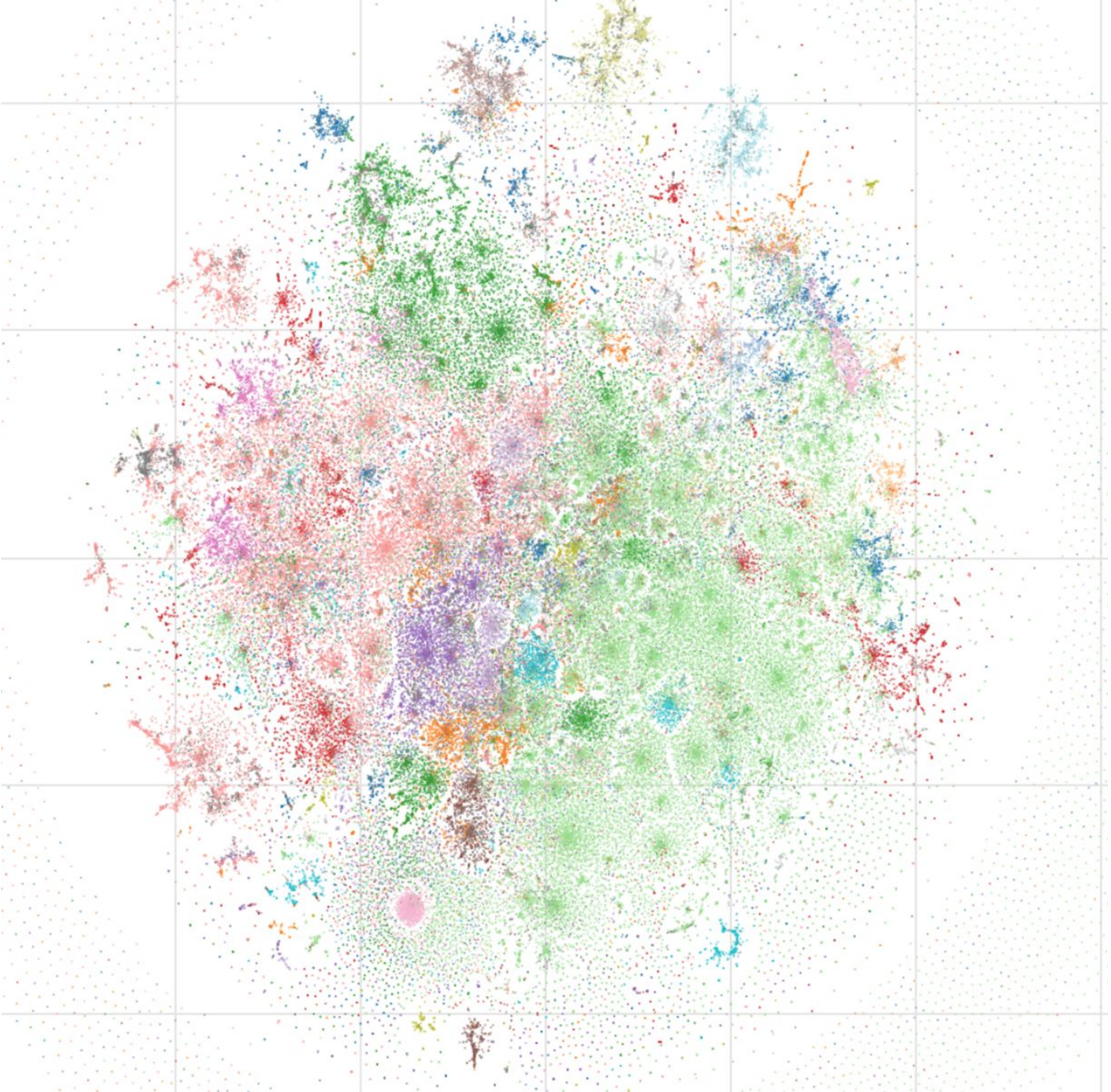
Towards comparison with reality

As comparison with reality is such an open ended problem we do not have one individual exercise that proves for certain that we have captured reality.

We can, however, provide some exercises through which it can be confirmed that, at the least, doc2vec is producing reasonable results.

Reflection of Reality





Outline

Background and Motivation
- word2vec/doc2vec

Barriers to application
- Statistical robustness
- Descriptive power
- Reflect reality?

Our work
- Quantifying robustness
- Quantifying descriptive power
- Towards comparisons with reality

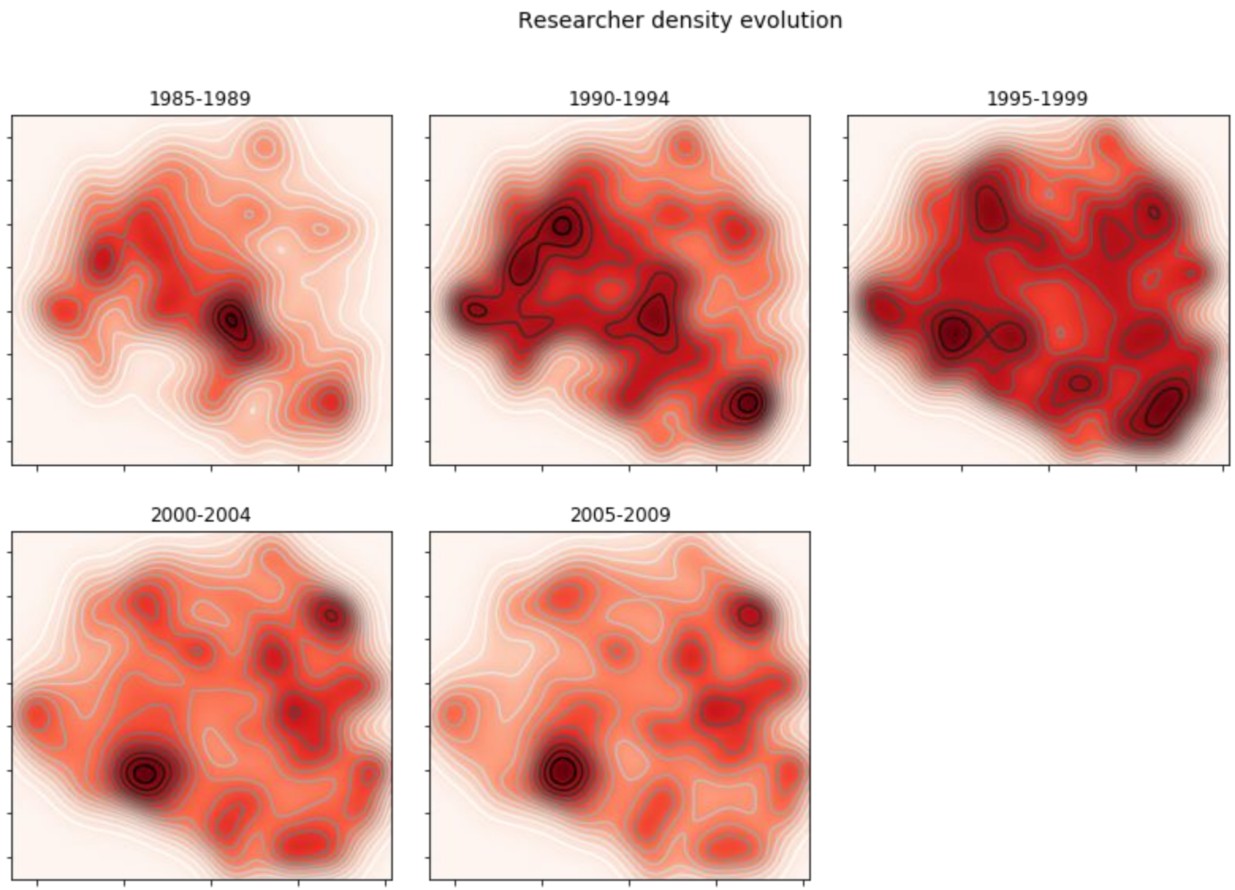
Wrap up

Wrap up

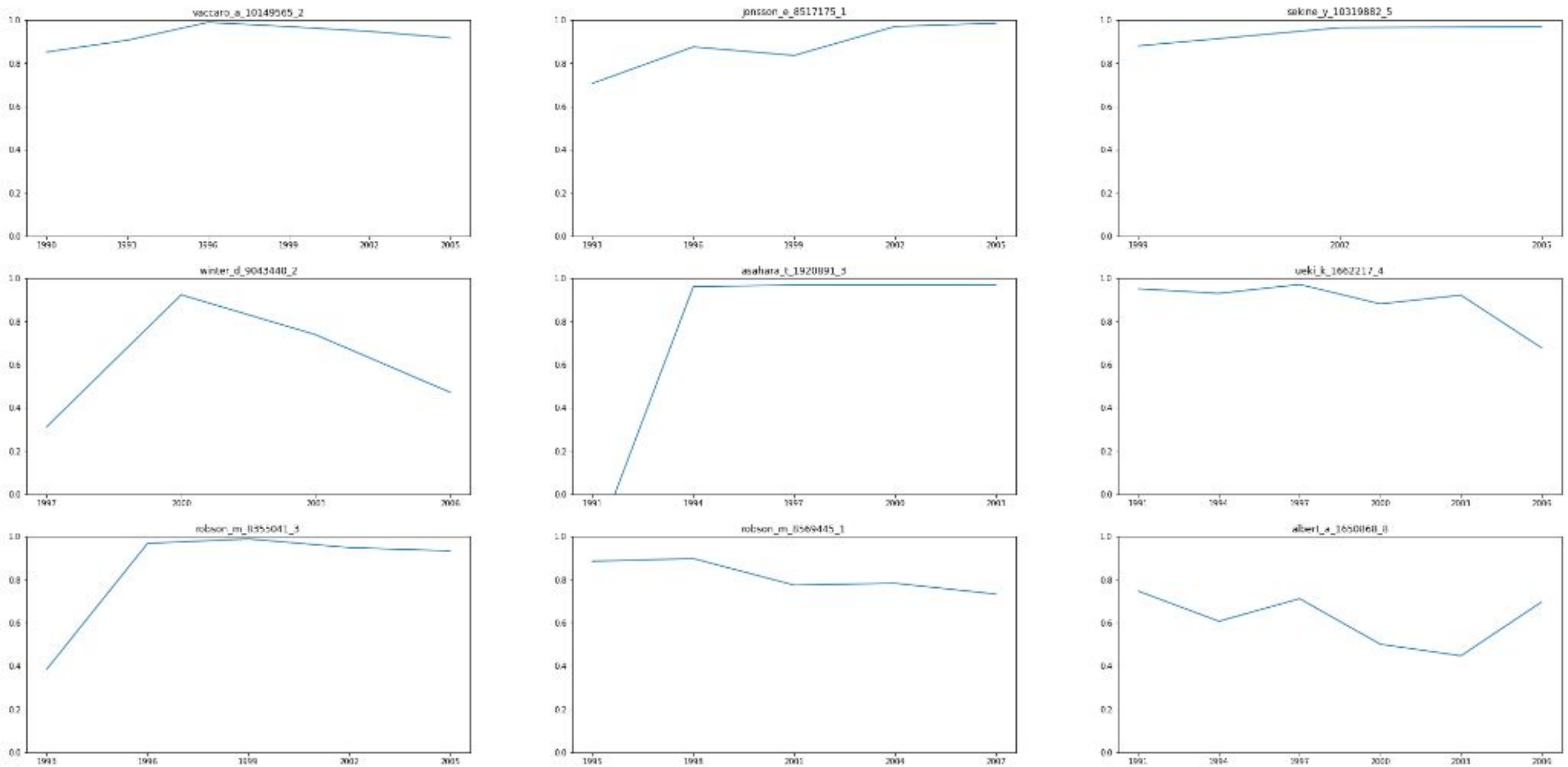
We develop and apply a methodology for evaluating the statistical robustness of topic models.

And in doing so, we find that the neural-network based doc2vec produces the best results.

Future Work


Researcher density evolution
1985-1989
1990-1994
1995-1999
2000-2004
2005-2009

Future Work



Robust similarity measures from topic modeling: validation and use.
9th Global TechMining Conference
Atlanta 17.10.2019
O Ballester*, O. Penner

Thanks

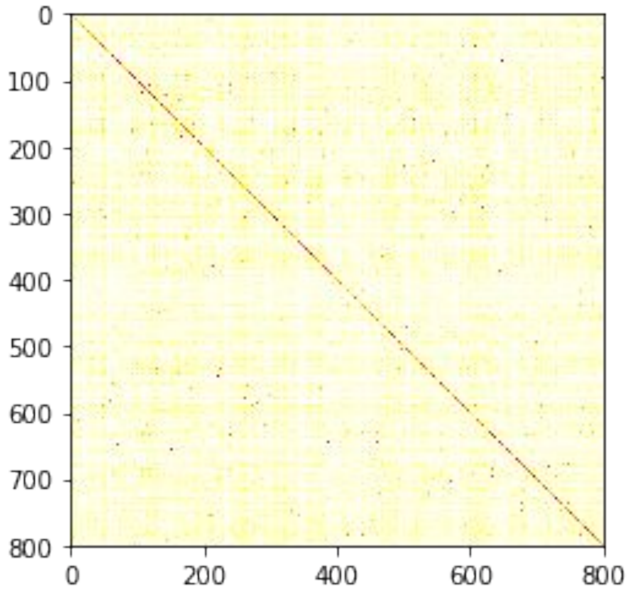
Robust similarity measures from topic modeling: validation and use

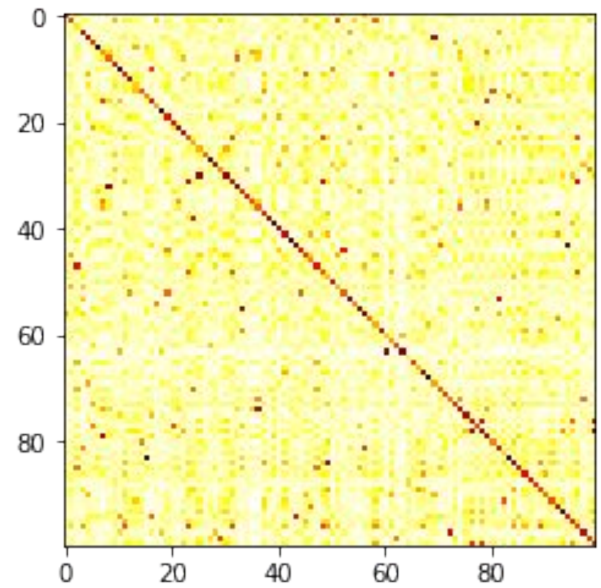
9th Global TechMining Conference

Atlanta 17.10.2019

O Ballester*, O. Penner

Statistical Robustness





Steyvers, Griffiths (2004, 2013)

Economics of Science

- Similarity between publications is of longstanding interest (Azoulay 2010, Furman 2015, Uzzi 2015…)

- Topic Classification goes even further back… started with Garfield (1965). He identified a goal of an "association-of-ideas" index.

- Most research has used ad-hoc measures using for contextual similarity:
    - Citation (or co-citation)
    - Combination/co-occurrence of (key)words
    - Given Classifications
    - Network measures

Word2Vec & Doc2Vec

We can argue they are not *technically* topic models

- Classifier predicting “missing” word
- Output is a non-sparse embedding…
- Topic-space (latent-space) has no direct nor easy interpretation...

$$
\begin{array}{cc}
\begin{array}{|c|c|c|c|}
\hline
.5 & .9 & -.4 & \cdots \\
\hline
.1 & -.3 & .1 & \cdots \\
\hline
-.7 & .2 & -.1 & \cdots \\
\hline
\vdots & \vdots & \vdots & \ddots \\
\hline
\end{array}
&
\begin{array}{c}
\text{W 1} \\
\text{W 2} \\
\text{D 3} \\

\end{array}
\end{array}
$$