A Statistical Approach for Quantifying Group Difference in Topic Distributions Using Clinical Discourse Samples

Grace O. Lawley[1], Peter A. Heeman[1], Jill K. Dolata[2], Eric Fombonne[3], Steven Bedrick[4]

[1]Computer Science and Engineering
[2]Department of Pediatrics
[3]Department of Psychiatry
[4]Department of Medical Informatics and Clinical Epidemiology
Oregon Health & Science University, Portland, Oregon, USA

SIGdial & INLG 2023
September 13th, 2023

OHSU

Outline of this presentation

1. Motivation
2. Describe a statistical approach for explore and quantify topic distributions captured by topic models
3. Demonstrate its application using LDA and 2 corpora
    - 20Newsgroups — Usenet posts from different topics
    - Clinical corpus — Language samples of Autistic* and Typically Developing (TD) children

* We are using identity-first language (i.e., Autistic children) here instead of person-first language (i.e., children with Autism) as the former is the current preference among many Autistic individuals (Brown, n.d.).

Motivation

- Topic modeling
  - Many different topics covered over course of a text or dialogue
  - Grouping documents into categories of topics covered

Motivation

- Topic modeling
  - Many different topics covered over course of a text or dialogue
  - Grouping documents into categories of topics covered

- Current methods for evaluating topic distributions
  - Intrinsic methods, such as within-topic coherence
  - To our knowledge, shortage of methods for *statistical* comparisons

Motivation

- Topic modeling
  - Many different topics covered over course of a text or dialogue
  - Grouping documents into categories of topics covered

- Current methods for evaluating topic distributions
  - Intrinsic methods, such as within-topic coherence
  - To our knowledge, shortage of methods for *statistical* comparisons

- Latent Dirichlet Allocation (LDA; Blei et al., 2003)
  - Capture and quantify topic distributions for a collection of language samples

Latent Dirichlet Allocation (LDA)

- LDA is a unsupervised, generative probabilistic model that is used on a corpus of text documents to model each document as a finite mixture over k topics
- Each document
  - Treated as a bag-of-words
  - Represented as a set of words and associated frequencies

Latent Dirichlet Allocation (LDA)

- LDA is a unsupervised, generative probabilistic model that is used on a corpus of text documents to model each document as a finite mixture over k topics
- Each document
  - Treated as a bag-of-words
  - Represented as a set of words and associated frequencies
- Given M documents and an integer k, LDA produces
  - $M \times k$ document-topic matrix (θ)
  - $k \times V$ topic-word matrix (β) — where V is total number of unique words across entire corpus

Latent Dirichlet Allocation (LDA)

- LDA is a unsupervised, generative probabilistic model that is used on a corpus of text documents to model each document as a finite mixture over k topics
- Each document
  - Treated as a bag-of-words
  - Represented as a set of words and associated frequencies
- Given M documents and an integer k, LDA produces
  - $M \times k$ document-topic matrix (θ)
  - $k \times V$ topic-word matrix (β) — where V is total number of unique words across entire corpus

Statistical Approach (1 of 3)

- Document-topic matrix, θ
  - Each row = single document
  - Each column = single topic
- The elements in θ are the estimated proportion of words in a document that were generated by a given topic

$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,k} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,k} \\ \theta_{3,1} & \theta_{3,2} & \cdots & \theta_{3,k} \\ & & \vdots & \\ \theta_{M,1} & \theta_{M,2} & \cdots & \theta_{M,k} \end{bmatrix}$$

Statistical Approach (1 of 3)

- Document-topic matrix, θ
  - Each row = single document
  - Each column = single topic
- The elements in θ are the estimated proportion of words in a document that were generated by a given topic
- Each document can be represented as a k-dimensional topic distribution vector

$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,k} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,k} \\ \theta_{3,1} & \theta_{3,2} & \cdots & \theta_{3,k} \\ & & \vdots & \\ \theta_{M,1} & \theta_{M,2} & \cdots & \theta_{M,k} \end{bmatrix}$$

Statistical Approach (1 of 3)

- Document-topic matrix, θ
  - Each row = single document
  - Each column = single topic
- The elements in θ are the estimated proportion of words in a document that were generated by a given topic
- Each document can be represented as a k-dimensional topic distribution vector

$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,k} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,k} \\ \theta_{3,1} & \theta_{3,2} & \cdots & \theta_{3,k} \\ & & \vdots & \\ \theta_{M,1} & \theta_{M,2} & \cdots & \theta_{M,k} \end{bmatrix}$$

Statistical Approach (1 of 3)

- Document-topic matrix, θ
  - Each row = single document
  - Each column = single topic
- The elements in θ are the estimated proportion of words in a document that were generated by a given topic
- Each document can be represented as a k-dimensional topic distribution vector

$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,k} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,k} \\ \theta_{3,1} & \theta_{3,2} & \cdots & \theta_{3,k} \\ & & \vdots & \\ \theta_{M,1} & \theta_{M,2} & \cdots & \theta_{M,k} \end{bmatrix}$$

Statistical Approach (1 of 3)

- Document-topic matrix, θ
  - Each row = single document
  - Each column = single topic
- The elements in θ are the estimated proportion of words in a document that were generated by a given topic
- Each document can be represented as a k-dimensional topic distribution vector

$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,k} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,k} \\ \theta_{3,1} & \theta_{3,2} & \cdots & \theta_{3,k} \\ & & \vdots & \\ \theta_{M,1} & \theta_{M,2} & \cdots & \theta_{M,k} \end{bmatrix}$$

Statistical Approach (1 of 3)

- Document-topic matrix, θ
  - Each row = single document
  - Each column = single topic
- The elements in θ are the estimated proportion of words in a document that were generated by a given topic
- Each document can be represented as a k-dimensional topic distribution vector

$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,k} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,k} \\ \theta_{3,1} & \theta_{3,2} & \cdots & \theta_{3,k} \\ & & \vdots & \\ \theta_{M,1} & \theta_{M,2} & \cdots & \theta_{M,k} \end{bmatrix}$$

Statistical Approach (1 of 3)

- Document-topic matrix, θ
  - Each row = single document
  - Each column = single topic
- The elements in θ are the estimated proportion of words in a document that were generated by a given topic

$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,k} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,k} \\ \theta_{3,1} & \theta_{3,2} & \cdots & \theta_{3,k} \\ & & \vdots & \\ \theta_{M,1} & \theta_{M,2} & \cdots & \theta_{M,k} \end{bmatrix}$$

- Each document can be represented as a k-dimensional topic distribution vector
  - Feature vectors for document classification or clustering
  - Proxy for document content for qualitative analyses

Statistical Approach (1 of 3)

- Document-topic matrix, θ
  - Each row = single document
  - Each column = single topic
- The elements in θ are the estimated proportion of words in a document that were generated by a given topic

$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,k} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,k} \\ \theta_{3,1} & \theta_{3,2} & \cdots & \theta_{3,k} \\ & & \vdots & \\ \theta_{M,1} & \theta_{M,2} & \cdots & \theta_{M,k} \end{bmatrix}$$

- Each document can be represented as a k-dimensional topic distribution vector
  - Feature vectors for document classification or clustering
  - Proxy for document content for qualitative analyses

- *To our knowledge, a statistical method for comparing topic distribution vectors between groups of documents has not yet been proposed*

Statistical Approach (2 of 3)

- One reason for this is due to certain numerical properties of topic distribution vectors which make them unsuitable for many parametric statistical methods
    - Each component is bounded between 0 and 1
    - All components sum to 1

Statistical Approach (2 of 3)

- One reason for this is due to certain numerical properties of topic distribution vectors which make them unsuitable for many parametric statistical methods
  - Each component is bounded between 0 and 1
  - All components sum to 1
- Realized that topic distribution vectors meet the definition of compositional data since components are proportions and all sum to 1
- Compositional data (Aitchison 1982) are vectors of positive numbers that together represent parts of some whole
  - e.g., the demographic profile of a city, the mineral composition of rocks

Statistical Approach (2 of 3)

- One reason for this is due to certain numerical properties of topic distribution vectors which make them unsuitable for many parametric statistical methods
    - Each component is bounded between 0 and 1
    - All components sum to 1
- Realized that topic distribution vectors meet the definition of compositional data since components are proportions and all sum to 1
- Compositional data (Aitchison 1982) are vectors of positive numbers that together represent parts of some whole
    - e.g., the demographic profile of a city, the mineral composition of rocks
- Isometric logratio (ILR) transformation (Egozcue et al., 2003)
    - ILR: $S^D \rightarrow \mathbb{R}^{D-1}$
    - Maps compositional data from its original sample space (D-part simplex) into real space ($D-1$ Euclidean space) with all metric properties preserved
    - After the transformation, we are able to use classical multivariate analysis tools

Statistical Approach (3 of 3)

- Multivariate Analysis of Means (MANOVA)
  - Compares multivariate sample means
  - Requires a number of statistical assumptions to be met before using (described in more detail in the paper)
  - Examines effect of one discrete, independent variable on multiple dependent variables
    - Independent variable —> topic label // diagnostic group
    - Dependent variables —> topic distribution probabilities in the document-topic distribution matrix created by LDA, $\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,k-1}$ where $i = 1,2,\ldots,M$

Statistical Approach (3 of 3)

- Multivariate Analysis of Means (MANOVA)
  - Compares multivariate sample means
  - Requires a number of statistical assumptions to be met before using (described in more detail in the paper)
  - Examines effect of one discrete, independent variable on multiple dependent variables
    - Independent variable —> topic label // diagnostic group
    - Dependent variables —> topic distribution probabilities in the document-topic distribution matrix created by LDA, $\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,k-1}$ where $i = 1,2,\ldots,M$

One dimension is removed during the ILR transformation

Statistical Approach (3 of 3)

- Multivariate Analysis of Means (MANOVA)
  - Compares multivariate sample means
  - Requires a number of statistical assumptions to be met before using (described in more detail in the paper)
  - Examines effect of one discrete, independent variable on multiple dependent variables
    - Independent variable —> topic label // diagnostic group
    - Dependent variables —> topic distribution probabilities in the document-topic distribution matrix created by LDA, $\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,k-1}$ where $i = 1,2,\ldots,M$

One dimension is removed during the ILR transformation

- After MANOVA, calculate effect size
  - Partial eta-squared (η^2)
  - What proportion of the variance of the linear combination of topics can be explained by the independent variable

Statistical Approach

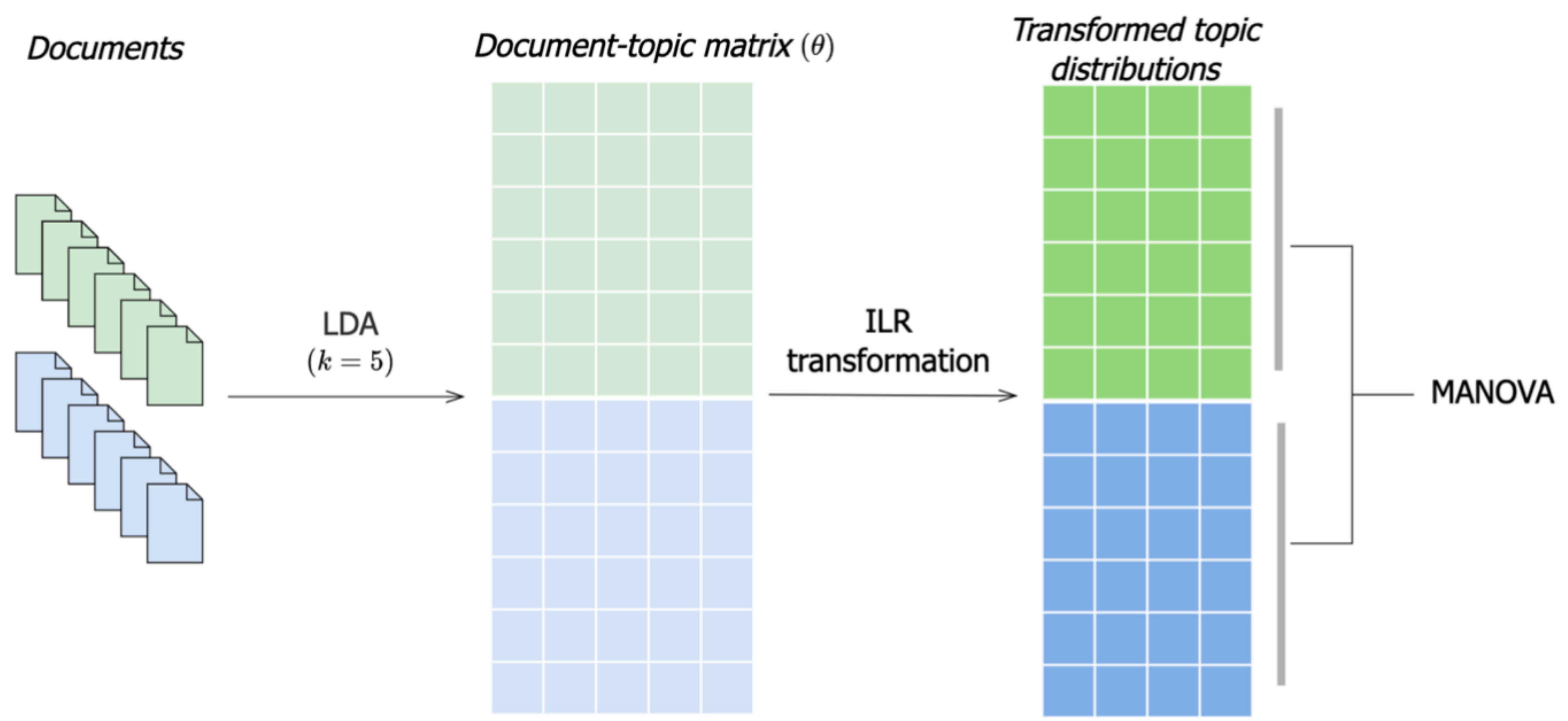


Figure 1: Example workflow for the described statistical approach described to explore and quantify group differences in topic distributions captured by topic models.

20NewsGroups (1 of 3)

- Collection of ~18,000 posts from twenty different Usenet* newsgroups
- Widely used for text classification and analysis

*Usenet was an early internet-based network of hierarchally-organized discussion groups where users could post messages about a given topic.

20NewsGroups (1 of 3)

- Collection of ~18,000 posts from twenty different Usenet* newsgroups
- Widely used for text classification and analysis
- Used documents from four topics
  - *comp.sys.ibm.pc.hardware*
  - *comp.sys.mac.hardware*
  - *rec.sport.baseball*
  - *rec.sport.hockey*

*Usenet was an early internet-based network of hierarchally-organized discussion groups where users could post messages about a given topic.

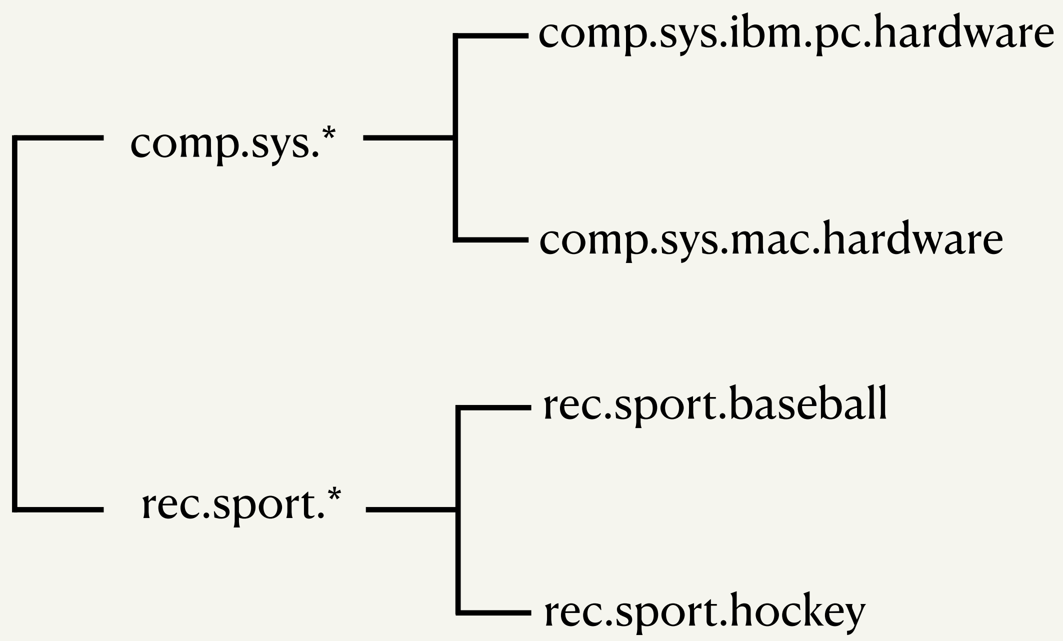
20NewsGroups (1 of 3)

- Collection of ~18,000 posts from twenty different Usenet* newsgroups
- Widely used for text classification and analysis
- Used documents from four topics
  - *comp.sys.ibm.pc.hardware*
  - *comp.sys.mac.hardware*
  - *rec.sport.baseball*
  - *rec.sport.hockey*

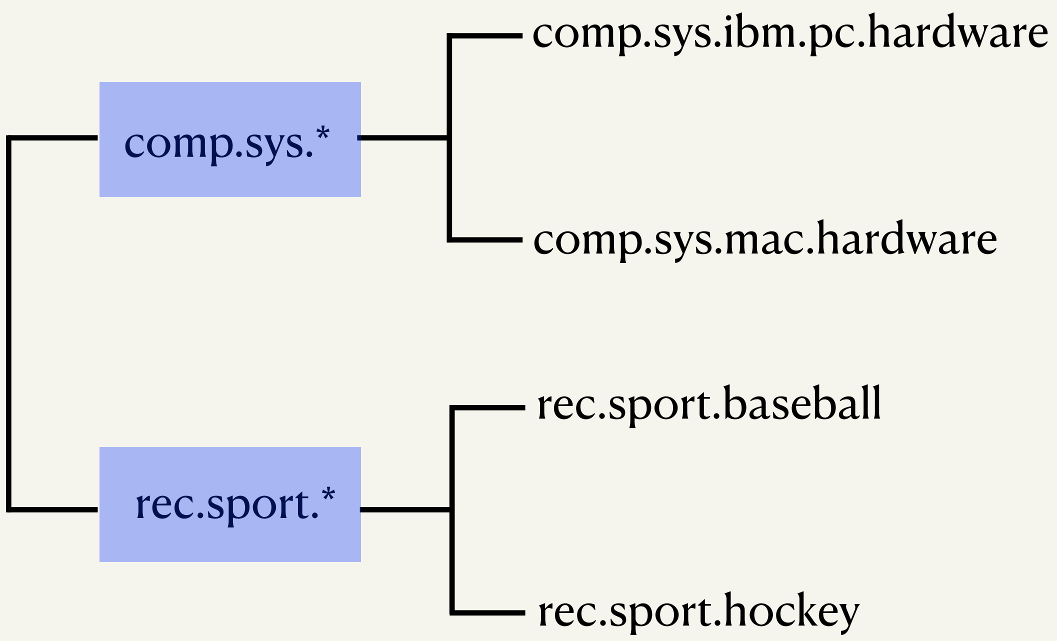
*Usenet was an early internet-based network of hierarchally-organized discussion groups where users could post messages about a given topic.

- Fit a single LDA model with a k value of 20
- Transformed topic distribution vectors using ILR transformation
- Checked MANOVA assumptions (detailed in paper)
- Performed 7 MANOVA tests

20NewsGroups (2 of 3)



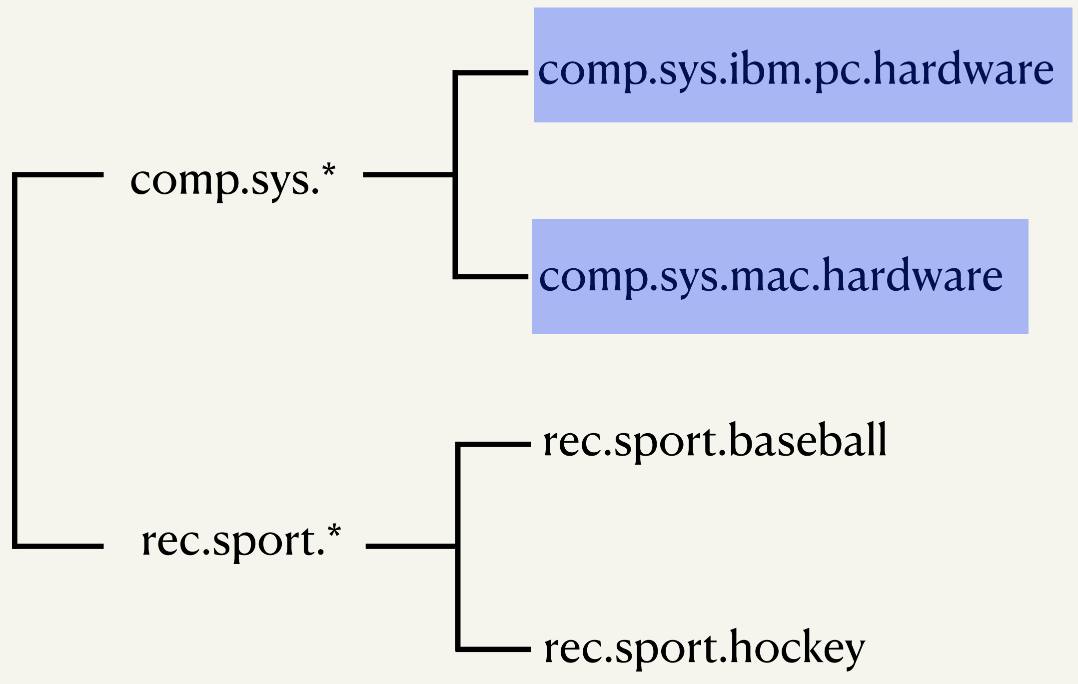
20NewsGroups (2 of 3)



1. Between broader categories (x1)

   Hypothesis: topic distributions will be very different

20NewsGroups (2 of 3)



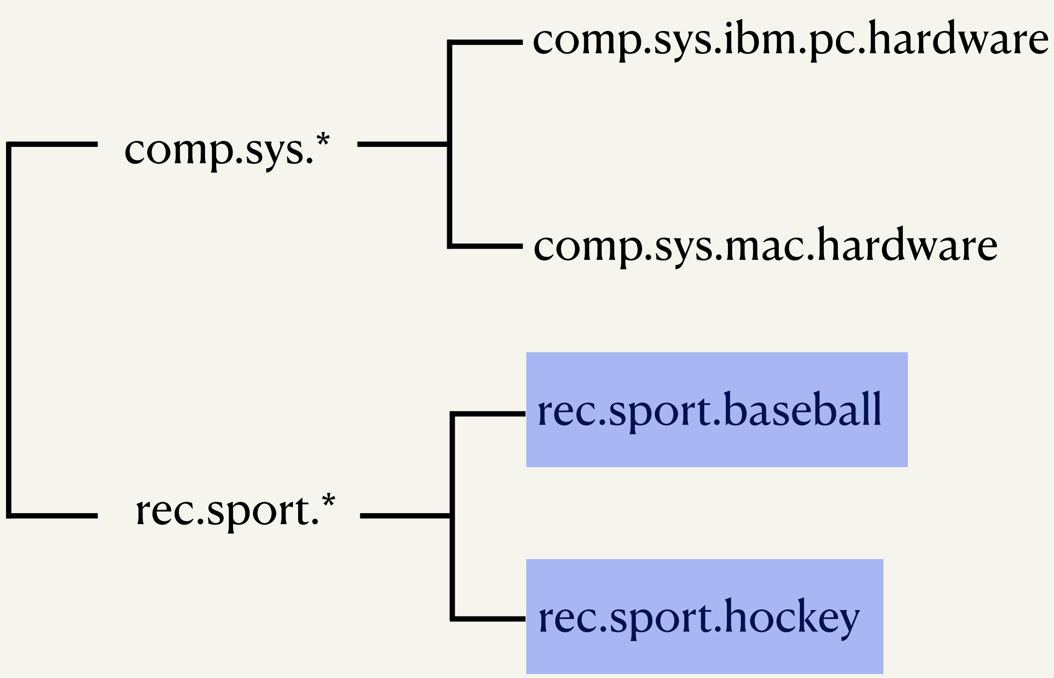
1. Between broader categories (x1)

   Hypothesis: topic distributions will be very different

2. Between subcategories (x2)

   Hypothesis: topic distributions will also be different, but not as different as previous comparison

20NewsGroups (2 of 3)



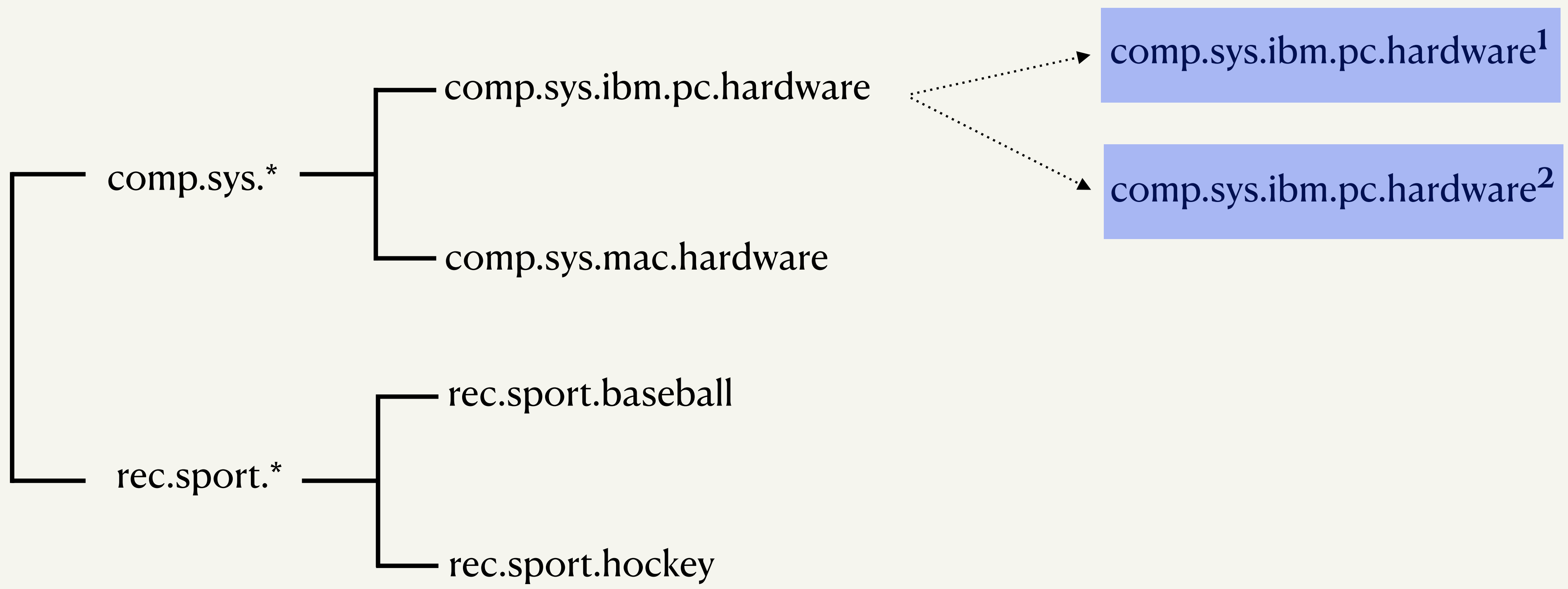
1. Between broader categories (x1)

   Hypothesis: topic distributions will be very different

2. Between subcategories (x2)

   Hypothesis: topic distributions will also be different, but not as different as previous comparison

20NewsGroups (2 of 3)



1. Between broader categories (x1)

   Hypothesis: topic distributions will be very different

2. Between subcategories (x2)

   Hypothesis: topic distributions will also be different, but not as different as previous comparison

3. Within a single topic (x4)

   Hypothesis: no difference between topic distributions

20NewsGroups (3 of 3)

1. Between broader categories

topics	n	df	Pillai	approx. F	df_1	df_2	p	partial η^2
*comp.sys.**	815	1	0.822	414.240	19	1710	<0.001	0.82
*rec.sport.**	915							

Table 2: *20Newsgroups*, comparison of LDA topic distribution vectors between and within topics.

20NewsGroups (3 of 3)

1. Between broader categories

2. Between subcategories

topics	n	df	Pillai	approx. F	df_1	df_2	p	partial η^2
*comp.sys.**	815	1	0.822	414.240	19	1710	<0.001	0.82
*rec.sport.**	915							
comp.sys.ibm.pc.hardware	447	1	0.264	15.008	19	795	<0.001	0.26
comp.sys.mac.hardware	368							
rec.sport.baseball	423	1	0.571	62.722	19	895	<0.001	0.57
rec.sport.hockey	492							

Table 2: *20Newsgroups*, comparison of LDA topic distribution vectors between and within topics.

20NewsGroups (3 of 3)

1. Between broader categories
2. Between subcategories

topics	n	df	Pillai	approx. F	df_1	df_2	p	partial η^2
*comp.sys.**	815	1	0.822	414.240	19	1710	<0.001	0.82
*rec.sport.**	915							
comp.sys.ibm.pc.hardware	447	1	0.264	15.008	19	795	<0.001	0.26
comp.sys.mac.hardware	368							
rec.sport.baseball	423	1	0.571	62.722	19	895	<0.001	0.57
rec.sport.hockey	492							

Table 2: *20Newsgroups*, comparison of LDA topic distribution vectors between and within topics.

20NewsGroups (3 of 3)

1. Between broader categories
2. Between subcategories
3. Within a single topic

topics	n	df	Pillai	approx. F	df_1	df_2	p	partial η^2
*comp.sys.**	815	1	0.822	414.240	19	1710	<0.001	0.82
*rec.sport.**	915							
comp.sys.ibm.pc.hardware	447	1	0.264	15.008	19	795	<0.001	0.26
comp.sys.mac.hardware	368							
rec.sport.baseball	423	1	0.571	62.722	19	895	<0.001	0.57
rec.sport.hockey	492							
comp.sys.ibm.pc.hardware	219	1	0.020	0.460	19	427	0.976	0.02
"	228							
comp.sys.mac.hardware	198	1	0.044	0.840	19	348	0.659	0.04
"	170							
rec.sport.baseball	206	1	0.041	0.903	19	403	0.579	0.04
"	217							
rec.sport.hockey	247	1	0.029	0.738	19	472	0.780	0.03
"	245							

Table 2: *20Newsgroups*, comparison of LDA topic distribution vectors between and within topics.

Clinical corpus (1 of 3)

- Autism Spectrum Disorder (ASD) is a developmental disorder
    - Social communication difficulties, such as problems with topic maintenance
- Sample of 117 ASD and 65 Typically Developing (TD) children, 4 to 15 years old
    - Transcribed dialogues between child and examiner during conversation activities in the ADOS

ADOS = Autism Diagnostic Observation Schedule

Clinical corpus (1 of 3)

- Autism Spectrum Disorder (ASD) is a developmental disorder
    - Social communication difficulties, such as problems with topic maintenance
- Sample of 117 ASD and 65 Typically Developing (TD) children, 4 to 15 years old
    - Transcribed dialogues between child and examiner during conversation activities in the ADOS
- Compare topic distributions in two ways, (1) within child speech (2) within examiner speech
    - For child speech, expect topic distribution vectors of ASD group to be different from those of their TD peers
    - For examiner speech, do not expect topic distributions to differ between ASD and TD groups

ADOS = Autism Diagnostic Observation Schedule

Clinical corpus (2 of 3)

- Fit two separate LDA models: one containing child speech and one containing examiner speech
- Document = all words said by a speaker during a single ADOS conversation activity
  - Four activity types —> each child-examiner conversation is associated with four, distinct documents

Clinical corpus (2 of 3)

- Fit two separate LDA models: one containing child speech and one containing examiner speech
- Document = all words said by a speaker during a single ADOS conversation activity
    - Four activity types —> each child-examiner conversation is associated with four, distinct documents
- k of 20 used for both models
    - Informed by prior knowledge of type and quantity of questions asked

Clinical corpus (2 of 3)

- Fit two separate LDA models: one containing child speech and one containing examiner speech
- Document = all words said by a speaker during a single ADOS conversation activity
  - Four activity types —> each child-examiner conversation is associated with four, distinct documents
- k of 20 used for both models
  - Informed by prior knowledge of type and quantity of questions asked
- MANOVA tests
  - Independent variable = diagnosis (ASD, TD)
  - Dependent variables = topic probability values from the document-topic vectors
  - Null hypothesis: multivariate means of ASD and TD groups are equal

Clinical corpus — Results

1. Child speech

		df	Pillai	approx. F	df_1	df_2	p	partial η^2
Emotions	dx	1	0.093	0.941	19	175	0.5334	0.09
Social	dx	1	0.188	2.055	19	169	0.0083	0.19
Friends	dx	1	0.131	1.388	19	175	0.1381	0.13
Loneliness	dx	1	0.135	1.275	19	156	0.207	0.13

Table 3: Child speech, comparison of LDA topic distribution vectors between ASD and TD groups.

2. Examiner speech

		df	Pillai	approx. F	df_1	df_2	p	partial η^2
Emotions	dx	1	0.195	2.235	19	175	0.0035	0.20
Social	dx	1	0.296	3.858	19	174	<0.001	0.30
Friends	dx	1	0.165	1.833	19	176	0.0224	0.17
Loneliness	dx	1	0.151	1.557	19	167	0.0726	0.15

Table 4: Examiner speech, comparison of LDA topic distribution vectors between ASD and TD groups.

Clinical corpus — Results

1. Child speech

		df	Pillai	approx. F	df_1	df_2	p	partial η^2
Emotions	dx	1	0.093	0.941	19	175	0.5334	0.09
Social	dx	1	0.188	2.055	19	169	0.0083	0.19
Friends	dx	1	0.131	1.388	19	175	0.1381	0.13
Loneliness	dx	1	0.135	1.275	19	156	0.207	0.13

Table 3: Child speech, comparison of LDA topic distribution vectors between ASD and TD groups.

2. Examiner speech

		df	Pillai	approx. F	df_1	df_2	p	partial η^2
Emotions	dx	1	0.195	2.235	19	175	0.0035	0.20
Social	dx	1	0.296	3.858	19	174	<0.001	0.30
Friends	dx	1	0.165	1.833	19	176	0.0224	0.17
Loneliness	dx	1	0.151	1.557	19	167	0.0726	0.15

Table 4: Examiner speech, comparison of LDA topic distribution vectors between ASD and TD groups.

Clinical corpus — Results

1. Child speech

		df	Pillai	approx. F	df_1	df_2	p	partial η^2
Emotions	dx	1	0.093	0.941	19	175	0.5334	0.09
Social	dx	1	0.188	2.055	19	169	0.0083	0.19
Friends	dx	1	0.131	1.388	19	175	0.1381	0.13
Loneliness	dx	1	0.135	1.275	19	156	0.207	0.13

Table 3: Child speech, comparison of LDA topic distribution vectors between ASD and TD groups.

2. Examiner speech

		df	Pillai	approx. F	df_1	df_2	p	partial η^2
Emotions	dx	1	0.195	2.235	19	175	0.0035	0.20
Social	dx	1	0.296	3.858	19	174	<0.001	0.30
Friends	dx	1	0.165	1.833	19	176	0.0224	0.17
Loneliness	dx	1	0.151	1.557	19	167	0.0726	0.15

Table 4: Examiner speech, comparison of LDA topic distribution vectors between ASD and TD groups.

Future work

- Approach is not restricted to LDA
  - Method can be extended to any topic modeling algorithm that outputs a topic distribution that can be treated as a composition and satisfies the assumption for MANOVA
- Could include additional independent variables by using multivariate analysis of covariance (MANCOVA)
  - For the clinical corpus, participant age, sex, and IQ

Thank you

A Statistical Approach for Quantifying Group Difference in Topic Distributions Using Clinical Discourse Samples

Grace O. Lawley, Peter A. Heeman, Jill K. Dolata, Eric Fombonne, Steven Bedrick

Github repo: https://github.com/gracelawley/lawley-sigdial-2023

I am expecting to graduate by the end of 2023 and am on the job market!

Grace Olive Lawley
PhD Candidate, Computer Science & Engineering
Oregon Health & Science University
Portland, Oregon, USA
https://grace.rbind.io

OHSU

This work was supported in part by the National Institute on Deafness and Other Communication Disorders of the NIH under Awards R01DC012033 (PI: Dr. E. Fombonne) and R01DC015999 (PIs: Dr. S. Bedrick & G. Fergadiotis).