Mixed Membership Markov Models for Unsupervised Conversation Modeling

MICHAEL J. PAUL
JOHNS HOPKINS UNIVERSITY

The Center For Language and Speech Processing
at the Johns Hopkins University

JOHNS HOPKINS
UNIVERSITY
WHITING SCHOOL OF ENGINEERING

Conversation Modeling: High Level Idea

- We'll be modeling sequences of documents
  - e.g. a sequence of email messages from a conversation

- We'll use M^4 = **M**ixed **M**embership **M**arkov **M**odels
- M^4 is a combination of
  - **Topic models (LDA, PLSA, etc.)**
    - Documents are mixtures of latent classes/topics
  - **Hidden Markov models**
    - Documents in a sequence depend on the previous document

Generative Models of Text

- Some distinctions to consider…

Inter-document structure

Intra-document structure

	Independent	**Markov**
Single-Class	Naïve Bayes	HMM
Mixed-Membership	LDA	This talk! ☺

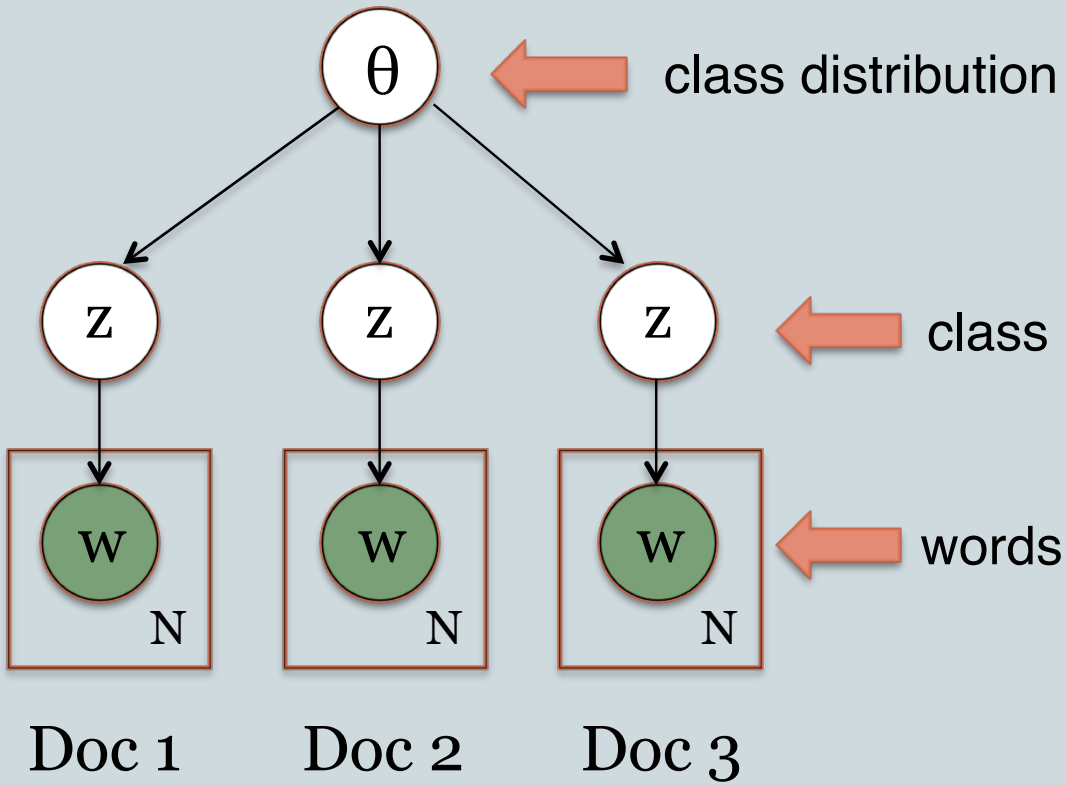
Overview

- Unsupervised Content Models
  - Naïve Bayes
  - Topic Models
- Unsupervised Conversation Modeling
  - Hidden Markov Models
- Mixed Membership Markov Models (M^4)
  - Overview
  - Inference
- Experiments with Conversation Data
  - Thread reconstruction
  - Speech act induction

Motivation: Unsupervised Models

- Huge amounts of unstructured and unannotated data on the Web
- Unsupervised models can help manage this data and are robust to variations in language and genre
- Tools like topic models can uncover interesting patterns in large corpora

(Unsupervised) Naïve Bayes



- Each document belongs to some category/class z
- Each class z is associated with its own distribution over words

(Unsupervised) Naïve Bayes

football	0.03
team	0.01
hockey	0.01
baseball	0.005
…	…

charge	0.02
court	0.02
police	0.015
robbery	0.01
…	…

congress	0.02
president	0.02
election	0.015
senate	0.01
…	…

probability distributions over words

imaginary class labels

"SPORTS"

"CRIME"

"POLITICS"

(Unsupervised) Naïve Bayes

football	0.03
team	0.01
hockey	0.01
baseball	0.005
…	…

Spanish team honored by fans, royal family in Madrid
SI.com - 21 minutes ago
Spain's national team received a joyous welcome at a parade through the streets of Madrid after winning Euro 2012 over Italy. MADRID (AP) -- Swathed in the red-and-yellow colors of Spain, hundreds of thousands packed central Madrid to give a


Telegraph.c...

charge	0.02
court	0.02
police	0.015
robbery	0.01
…	…

$21M lawsuit filed in NY police shooting
Fox News - 1 hour ago
WHITE PLAINS, NY - Police in suburban New York responding to a medical alert used excessive force when they killed an emotionally disturbed 68-year-old ex-Marine, the man's son claimed in a $21 million lawsuit Monday.


New York D...

congress	0.02
president	0.02
election	0.015
senate	0.01
…	…

Voters encouraged to use 'MyVote' before primary
Auburn Reporter - 1 hour ago
Many voters want a quick and easy way to learn more about the candidates they'll see on their primary ballot. Others simply want a fast and convenient way to register to vote or update their registration status in time for the primary.

Reporter Ne...

(Unsupervised) Naïve Bayes?

football	0.03
team	0.01
hockey	0.01
baseball	0.005
…	…

charge	0.02
court	0.02
police	0.015
robbery	0.01
…	…

congress	0.02
president	0.02
election	0.015
senate	0.01
…	…

What if an article belongs to more than one category?

(Unsupervised) Naïve Bayes?

football	0.03
team	0.01
hockey	0.01
baseball	0.005
…	…

charge	0.02
court	0.02
police	0.015
robbery	0.01
…	…

congress	0.02
president	0.02
election	0.015
senate	0.01
…	…

Jury Finds Baseball Star Roger Clemens Not Guilty On All Counts

A **jury** found **baseball** star **Roger Clemens** not **guilty** on six **charges** against. **Clemens** was **accused** of **lying** to **Congress** in 2008 about his use of **performance** enhancing **drugs**.

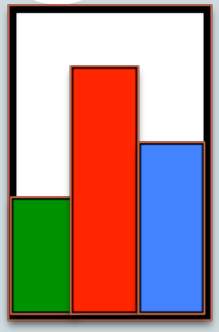
Topic Models

football	0.03
team	0.01
hockey	0.01
baseball	0.005
…	…

charge	0.02
court	0.02
police	0.015
robbery	0.01
…	…

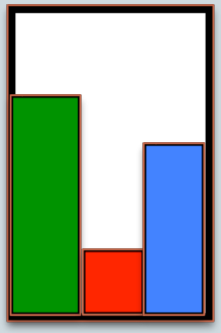
congress	0.02
president	0.02
election	0.015
senate	0.01
…	…

Doc 1

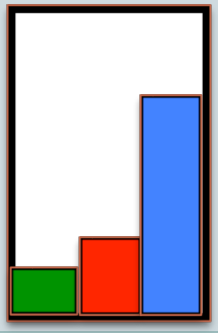


…

Doc 2

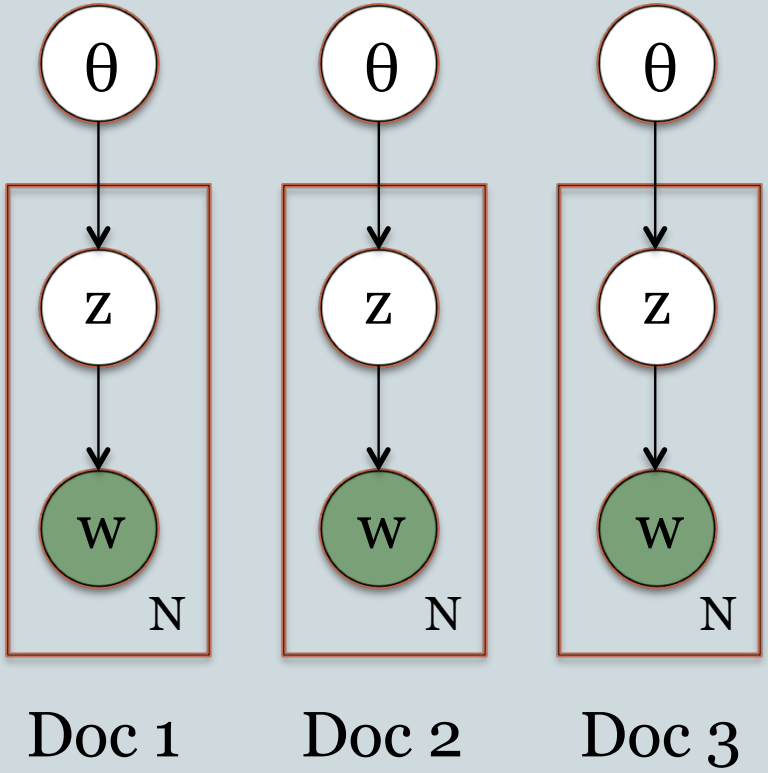


Doc 3



…

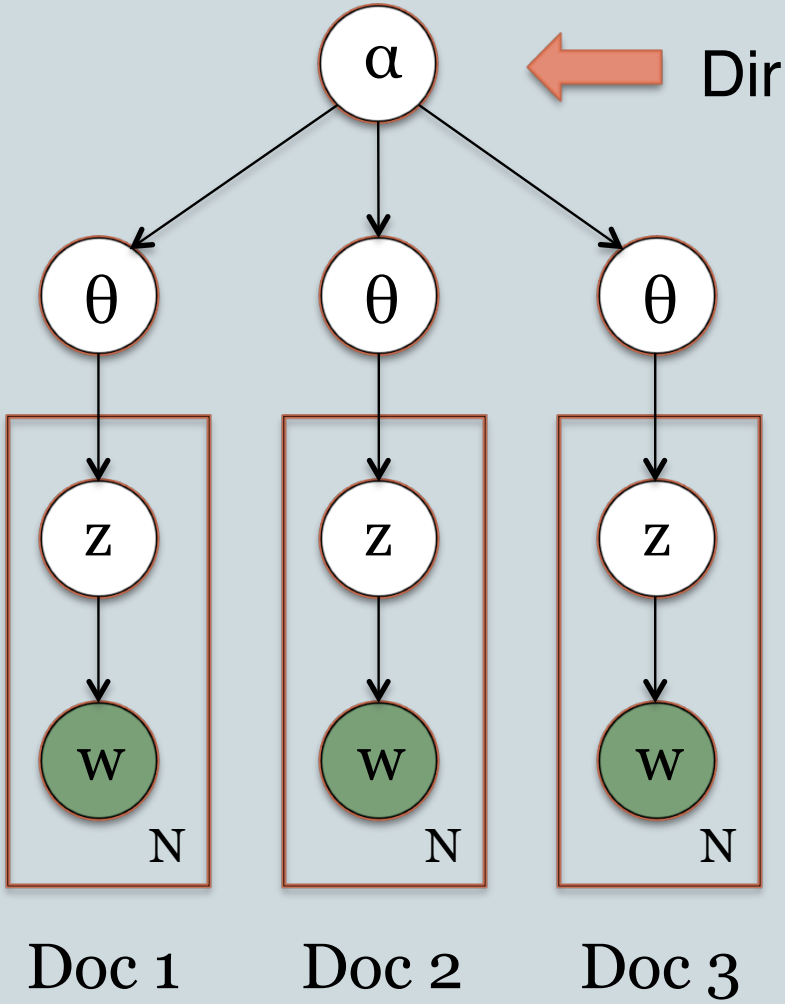
Topic Models



- One class distribution θ_d per document
- One class value per token
  - (rather than per document)

T. Hofmann. Probabilistic Latent Semantic Indexing. SIGIR 1999.

Latent Dirichlet Allocation (LDA)



Dirichlet prior

D. Blei, A. Ng, M. Jordan. Latent Dirichlet Allocation. JMLR 2003.

- One class distribution θ_d per document
- One class value per token
  - (rather than per document)

Overview

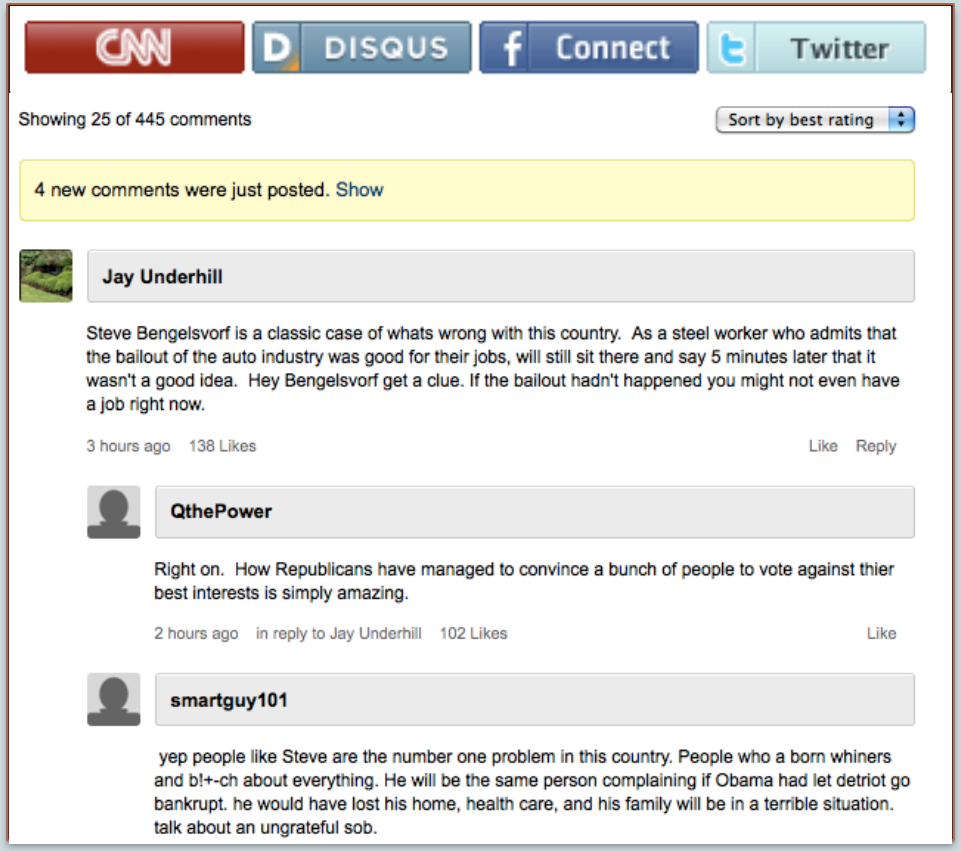
- Unsupervised Content Models
- **Unsupervised Conversation Modeling**
- Mixed Membership Markov Models
- Experiments with Conversation Data
- Conclusion

Conversation Modeling

- Documents on the web are more complicated than news articles

Conversation Modeling

- Documents on the web are more complicated than news articles

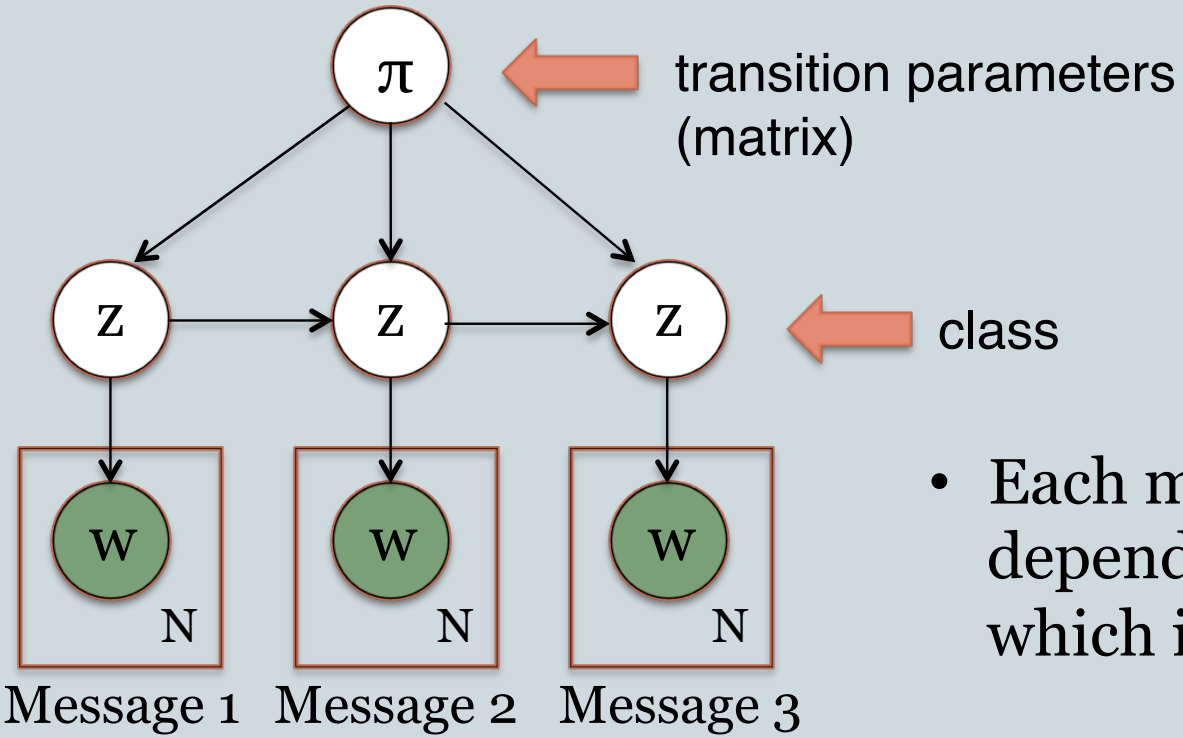


Conversation Modeling

- What's missing from Naïve Bayes and LDA?
  - They assume documents are generated independently of each other
- Messages in conversations aren't at all independent
  - Doesn't make sense to pretend that they are
  - But we'd like to represent this dependence in a reasonably simple way
- Solution: Hidden Markov Models

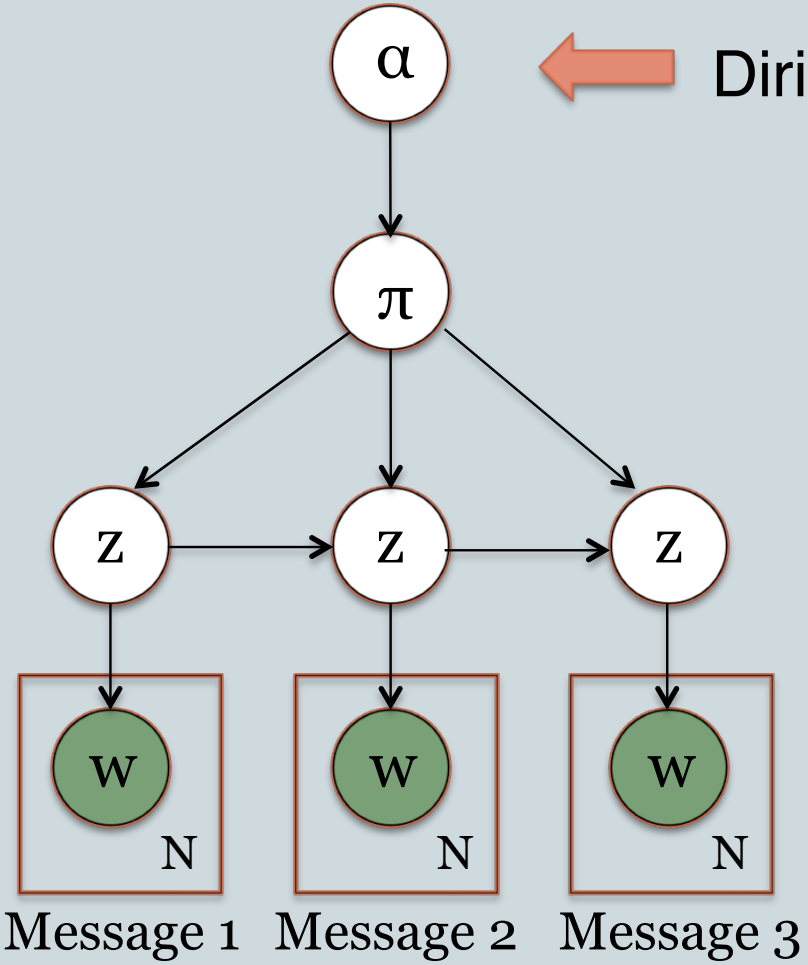
Block HMM

- Message emitted at each time step of Markov chain



- Each message in thread depends on the message to which it is a response

Bayesian Block HMM



Dirichlet prior

A. Ritter, C. Cherry, B. Dolan. Unsupervised Modeling of Twitter Conversations. HLT-NAACL 2010.

- Each message in thread depends on the message to which it is a response

Block HMM

hey	0.1
sup	0.06
hi	0.04
hello	0.01
…	…

GREETING

SPORTS

football	0.03
team	0.01
hockey	0.01
baseball	0.005
…	…

what	0.03
what's	0.025
how	0.02
is	0.02
…	…

QUESTION

CRIME

charge	0.02
court	0.02
police	0.015
robbery	0.01
…	…

lol	0.04
haha	0.04
:)	0.03
lmao	0.01
…	…

LAUGHTER

POLITICS

congress	0.02
president	0.02
election	0.015
senate	0.01
…	…

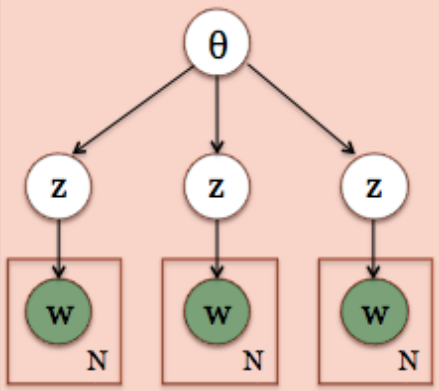
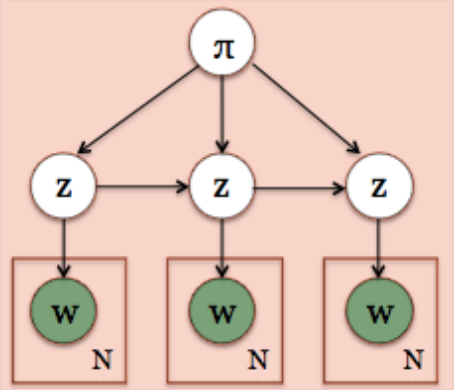
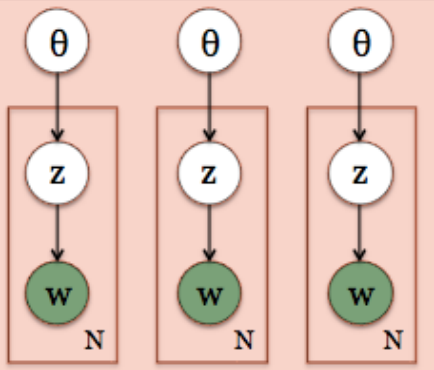
Block HMM

- Nice and simple way to model dependencies between messages
- This is similar to Naïve Bayes
  - One class per document!
- Let's make it more like LDA
  - Documents are *mixtures* of classes

Generative Models of Text

Inter-document structure

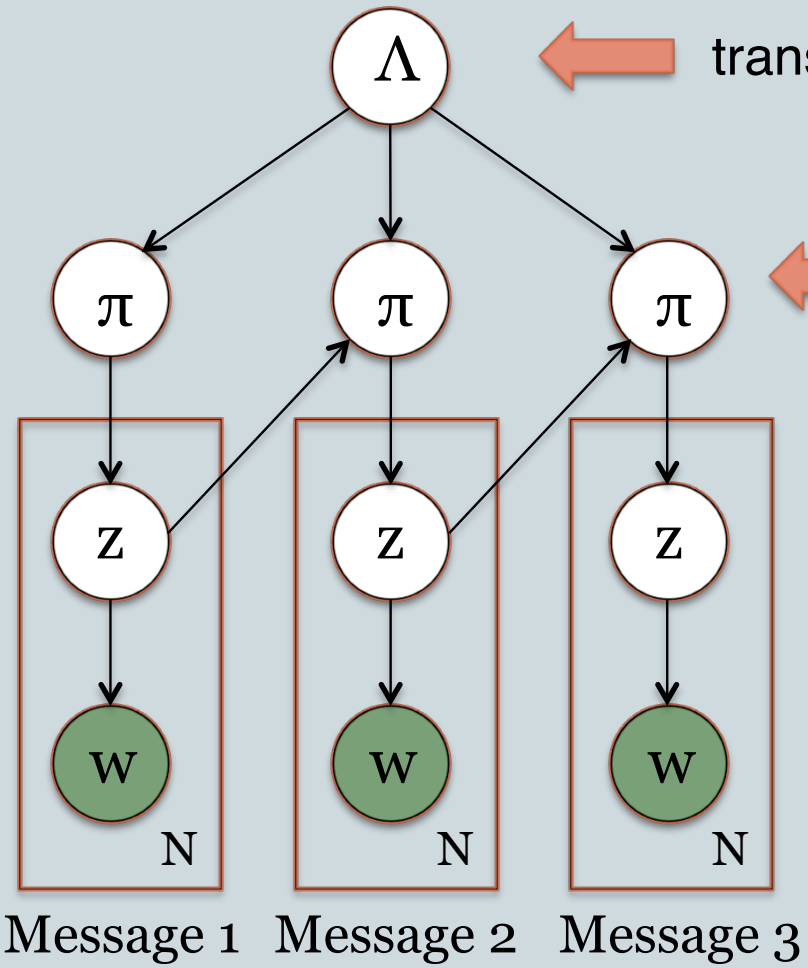
Intra-document structure

	Independent	**Markov**
Single-Class		
Mixed-Membership		This talk! ☺

Overview

- Unsupervised Content Models
- Unsupervised Conversation Modeling
- **Mixed Membership Markov Models**
- Experiments with Conversation Data
- Conclusion

Mixed Membership Markov Models (M^4)

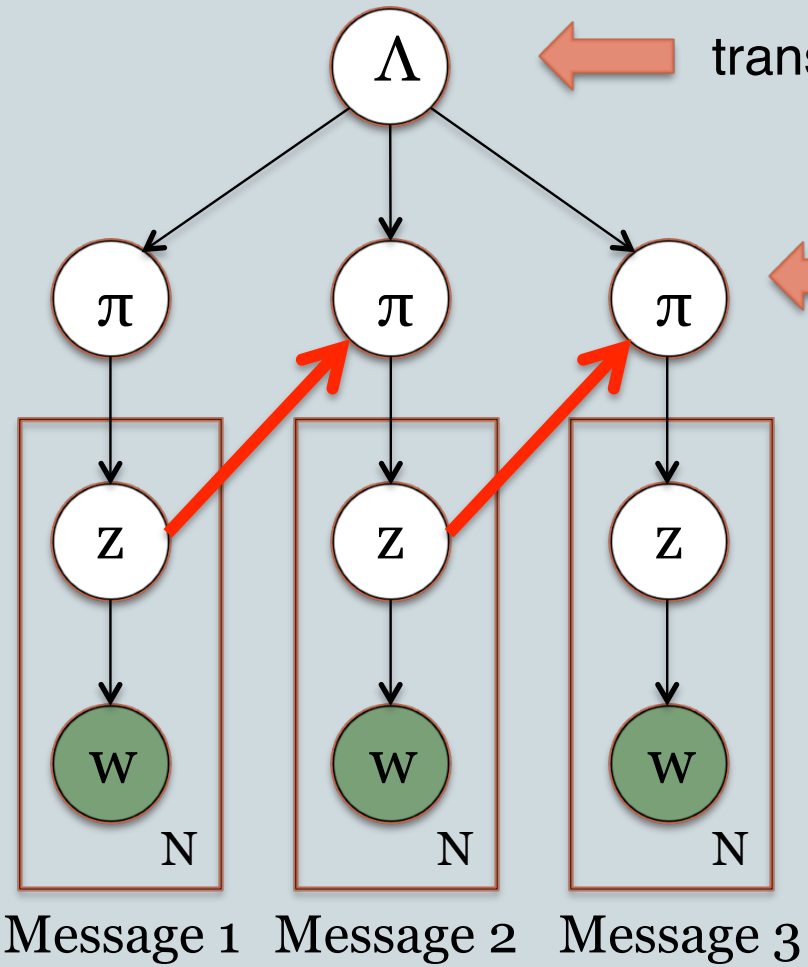


transition parameters

class distribution
(function of z and λ)

- Like LDA
  - One distribution π_d per doc
  - One class z per token
- But now each message's distribution depends on the class assignments of previous message

Mixed Membership Markov Models (M^4)



transition parameters

class distribution
(function of z and λ)

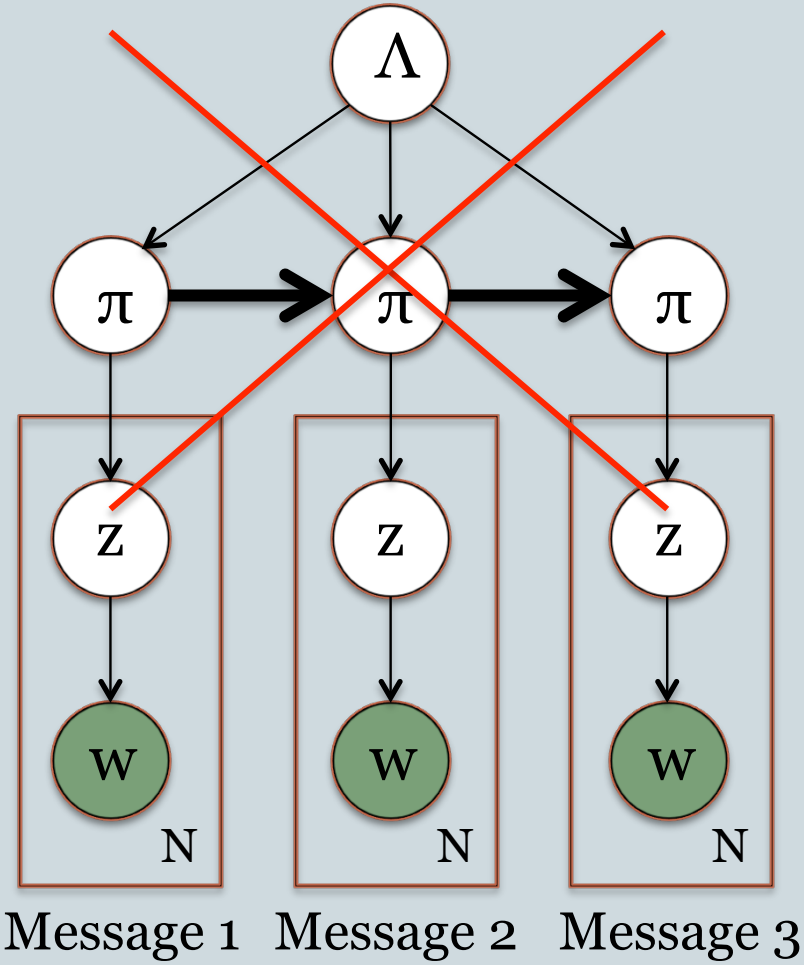
- **Core of M^4:**

Probability of class *j* in message *d*

$$\pi_{dj} \propto \exp(\lambda_j^T z_{d-1})$$

log-linear function
of previous message

Mixed Membership Markov Models (M^4)



- Why not transition directly from π to π?
- Makes more sense for next message to depend on actual classes of previous message (not the distribution over all possible classes)

Example

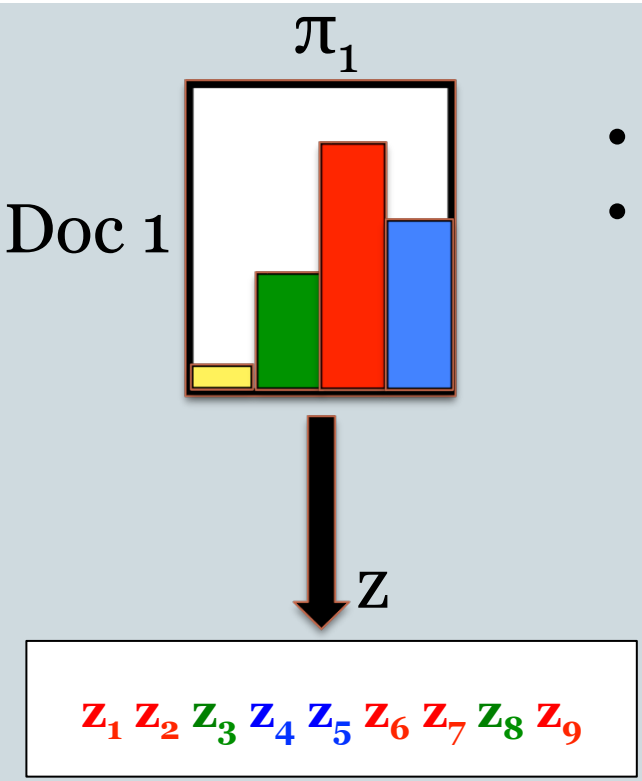
Suppose documents are mixtures of 4 classes: **Y** **G** **R** **B**

Then Λ is a 4x4 matrix with values such as:

$\lambda_{G \to R} = -0.2$ → "The presence of **G** in doc 1 slightly decreases the likelihood of having **R** in doc 2"

$\lambda_{B \to B} = 5.0$ → "The presence of **B** in doc 1 greatly increases the likelihood of having **B** in doc 2"

Example



- Multinomial parameters π
- Repeatedly sample z from π
  - i.e. sample class histogram for doc 1

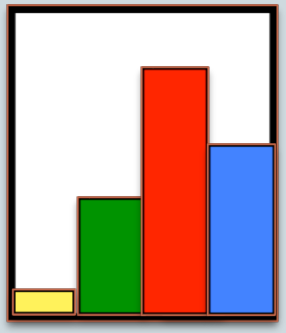
Counts of **z**:

Y: 0
G: 2
R: 5
B: 2

Example

π_1

Doc 1



Counts of **z**:

Y: 0
G: 2
R: 5
B: 2

$$\pi_{2Y} \propto \exp(\, 0\times\lambda_{Y\to Y} + 2\times\lambda_{G\to Y} + 5\times\lambda_{R\to Y} + 2\times\lambda_{B\to Y}\,) =$$

$$\pi_{2G} \propto \exp(\, 0\times\lambda_{Y\to G} + 2\times\lambda_{G\to G} + 5\times\lambda_{R\to G} + 2\times\lambda_{B\to G}\,) =$$

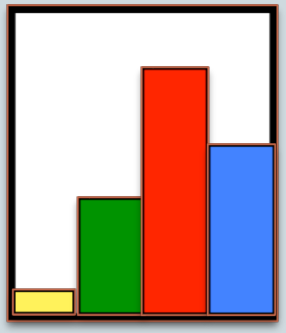
$$\pi_{2R} \propto \exp(\, 0\times\lambda_{Y\to R} + 2\times\lambda_{G\to R} + 5\times\lambda_{R\to R} + 2\times\lambda_{B\to R}\,) =$$

$$\pi_{2B} \propto \exp(\, 0\times\lambda_{Y\to B} + 2\times\lambda_{G\to B} + 5\times\lambda_{R\to B} + 2\times\lambda_{B\to B}\,) =$$

Example

π_1

Doc 1

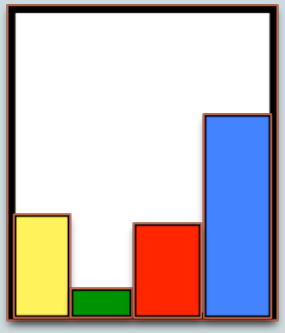


Counts of **z**:

Y: 0
G: 2
R: 5
B: 2

π_2

Doc 2



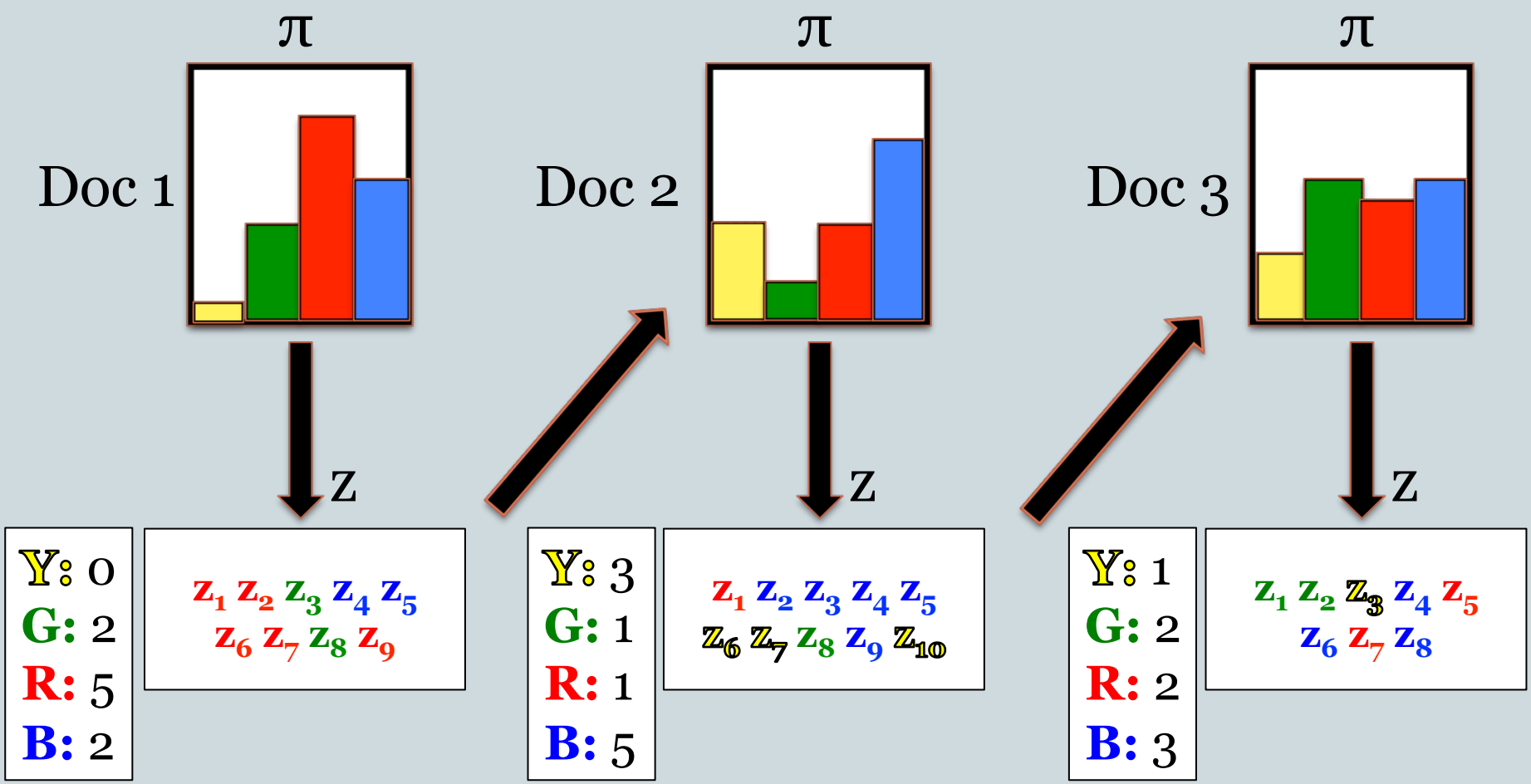
$$\pi_{2Y} \propto \exp(\, 0\times\lambda_{Y\to Y} + 2\times\lambda_{G\to Y} + 5\times\lambda_{R\to Y} + 2\times\lambda_{B\to Y}\,) =$$

$$\pi_{2G} \propto \exp(\, 0\times\lambda_{Y\to G} + 2\times\lambda_{G\to G} + 5\times\lambda_{R\to G} + 2\times\lambda_{B\to G}\,) =$$

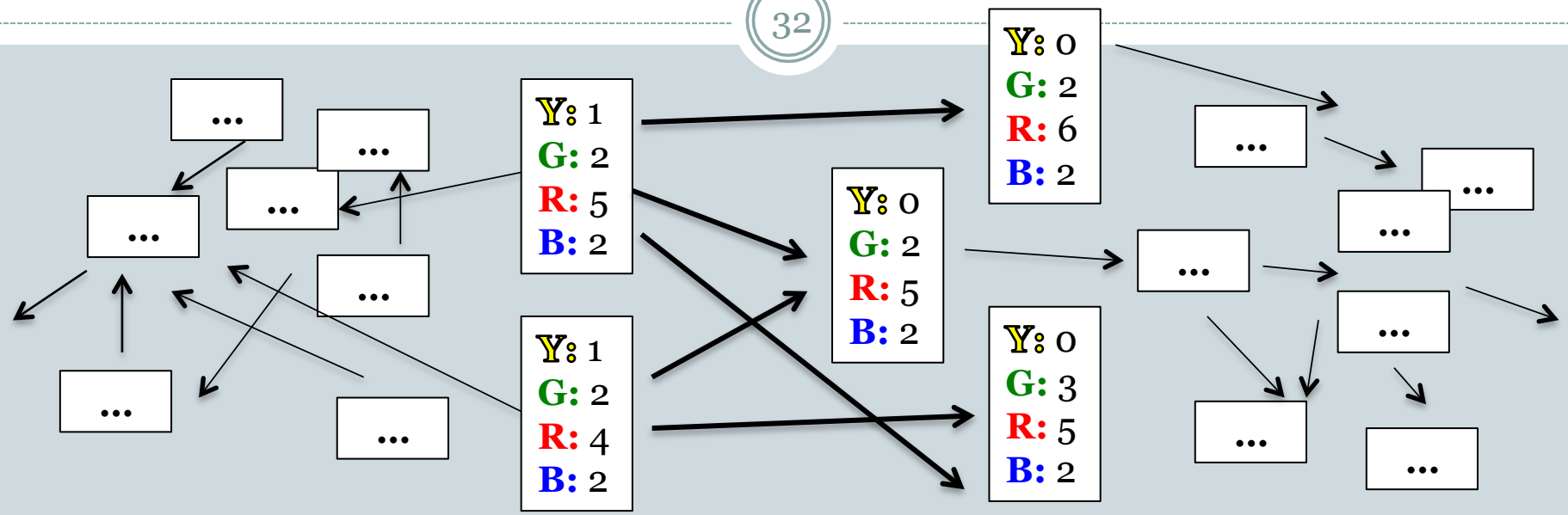
$$\pi_{2R} \propto \exp(\, 0\times\lambda_{Y\to R} + 2\times\lambda_{G\to R} + 5\times\lambda_{R\to R} + 2\times\lambda_{B\to R}\,) =$$

$$\pi_{2B} \propto \exp(\, 0\times\lambda_{Y\to B} + 2\times\lambda_{G\to B} + 5\times\lambda_{R\to B} + 2\times\lambda_{B\to B}\,) =$$

Example


π
Doc 1
z
Y: 0
G: 2
R: 5
B: 2
z_1 z_2 z_3 z_4 z_5
z_6 z_7 z_8 z_9
π
Doc 2
z
Y: 3
G: 1
R: 1
B: 5
z_1 z_2 z_3 z_4 z_5
z_6 z_7 z_8 z_9 z_{10}
π
Doc 3
z
Y: 1
G: 2
R: 2
B: 3
z_1 z_2 z_3 z_4 z_5
z_6 z_7 z_8

Mixed Membership Markov Models (M^4)



- M^4 is a Markov chain where the state space is the set of all possible class histograms
  - If no bound on document length, then the size of this space is countably infinite!
  - But the transition matrix is given in terms of the same number parameters as in a standard HMM

(Approximate) Inference

- Monte Carlo EM
  - E-step: Sample from posterior over class assignments (z)
  - M-step: Direct optimization of transition parameters (λ)
- Inference algorithm alternates between:
  - 1 iteration of collapsed Gibbs sampling
  - 1 iteration (step) of gradient ascent
- Sampler is similar to LDA Gibbs sampler
  - Slower because the computing the relative probability of each class involves summing over all classes to compute $\exp(\lambda_j^T z_{d-1})$

Overview

- Unsupervised Content Models
- Unsupervised Conversation Modeling
- Mixed Membership Markov Models
- **Experiments with Conversation Data**
- Conclusion

Data

- Two sets of asynchronous web conversations

- **CNET forums**
  - Technical help and discussion
  - Labeled with speech acts

S.N. Kim, L. Wang, T. Baldwin. Tagging and Linking Web Forum Posts. CoNLL 2010.

- **Twitter**
  - More personal communication
  - Short messages

# threads	# messages	# tokens per message
321	1309	78
36K	100K	13

Experimental Details

- Baselines:
  - Bayesian Block HMM (BHMM)
  - Latent Dirichlet Allocation (LDA)
- Symmetric Dirichlet prior on word distributions
  - Fancy way of describing smoothing
  - Concentration parameter sampled via Metropolis-Hastings
- 0-mean Gaussian prior on transition parameters λ
  - Independent weights (diagonal covariance)
  - Acts as L2 regularizer on weights
- All Dirichlet hyperparameters are optimized
  - Applies to LDA and BHMM

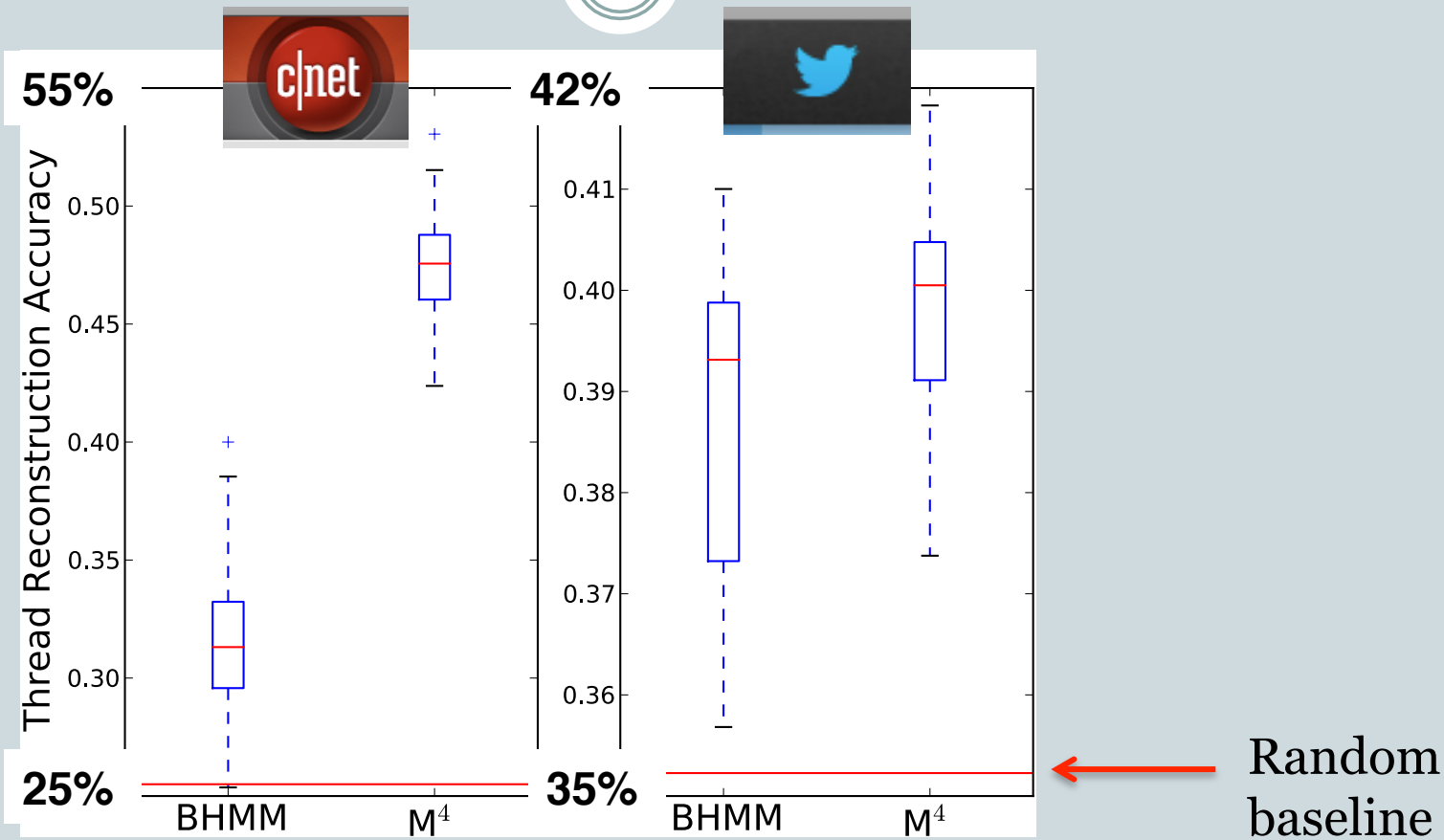
Thread Reconstruction

- Pretend we don't know the thread structure of a conversation. Can we figure out which messages are in response to which?

user1: hey **user2:** what's up? **user1:** not much	vs	**user1:** hey **user1:** not much **user2:** what's up?	vs	**user1:** not much **user2:** what's up? **user1:** hey

- Treat "parent" of each message as a hidden variable
  - Sample using simulated annealing
- Evaluate on held-out test data
  - Metric: **accuracy** (% of messages correctly aligned to parent)
  - Results pooled over many trials

Thread Reconstruction



- M4 is a lot better than Block HMM on CNET corpus
  - Twitter messages are short, so single-class assumption is probably reasonable

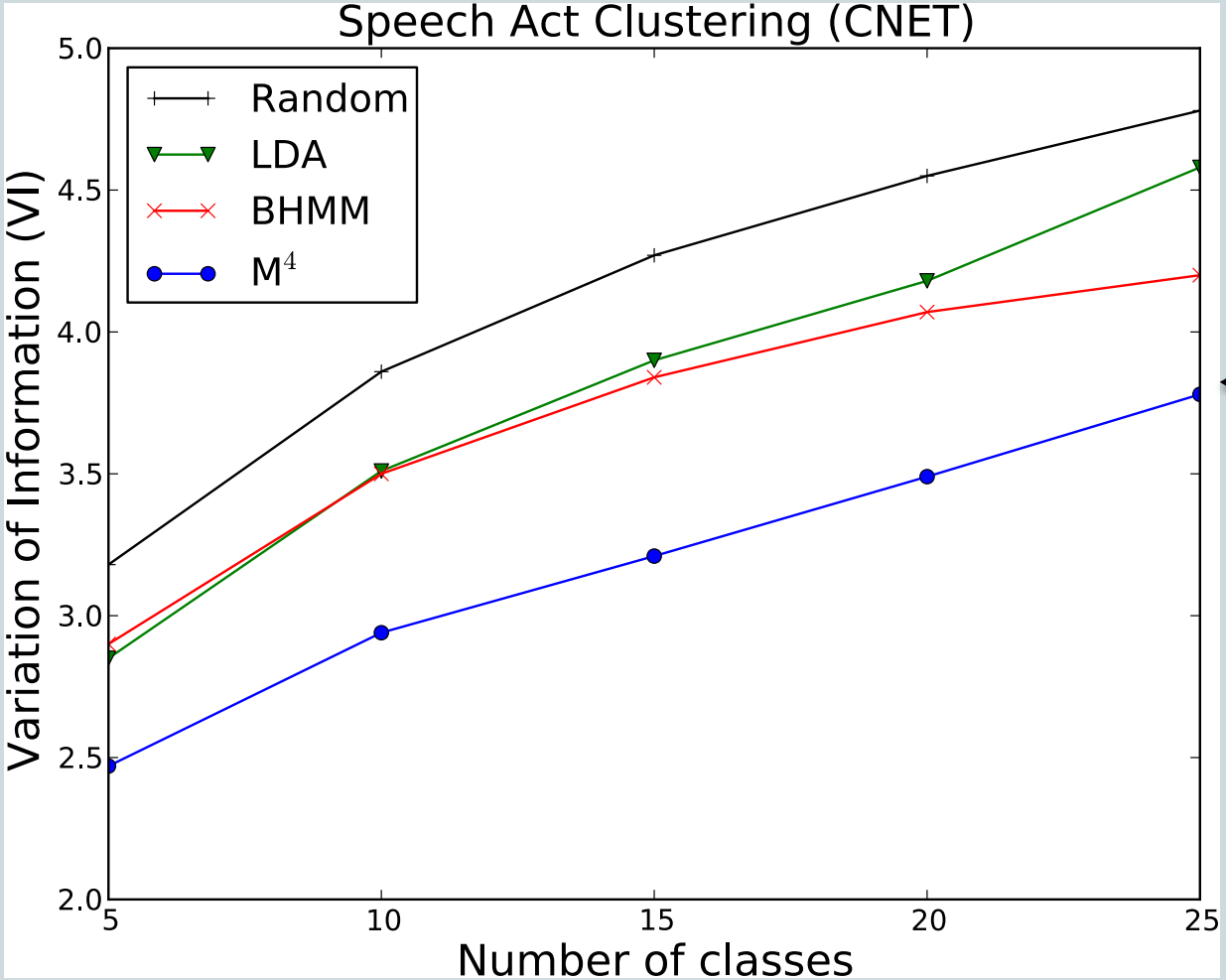
Speech Act Induction

- Messages in CNET corpus are annotated with speech act labels

- 12 labels
  - **Question** (broken into subclasses)
  - **Answer** (broken into subclasses)
  - **Resolution**, **Reproduction**, **Other**

- We measured how well the latent classes induced by M^4 matched the human labels
  - Metric: **variation of information (VI)**

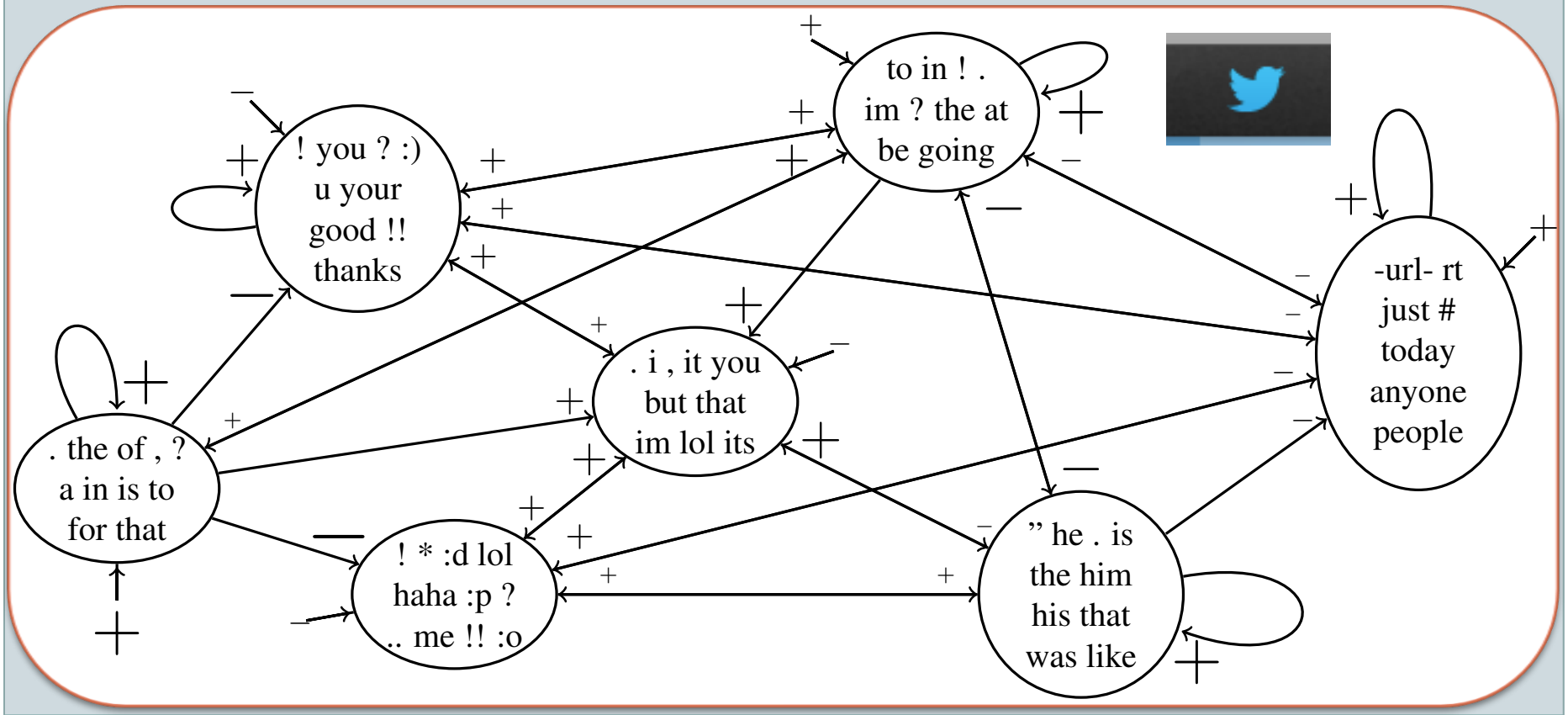
Speech Act Induction



M^4 is significantly better

What Does M4 Learn?

- Top words from a subset of classes
- Arrows show sign of λ from going from one class to another



Overview

- Unsupervised Content Models
- Unsupervised Conversation Modeling
- Mixed Membership Markov Models
- Experiments with Conversation Data
- **Conclusion**

Conclusion

- M^4
  - Combines properties of topic models and Markov models
  - Outperforms LDA and HMM individually
- Room for extensions
  - Richer model of **intra**-message structure
  - Bayesian formulations
- Code is available
  - `http://cs.jhu.edu/~mpaul`

Acknowledgements

- Advice:
  - Mark Dredze
  - Jason Eisner
  - Nick Andrews
  - Matt Gormley
  - Frank Ferraro, Wes Filardo, Adam Teichert, Tim Viera
- $$$:

NSF

JOHNS HOPKINS UNIVERSITY
WHITING SCHOOL OF ENGINEERING

Thank You 감사합니다

Perplexity

# classes:	5	10	15	20	25
CNET					
Unigram	63.1	63.1	63.1	63.1	63.1
LDA	57.2	54.4	52.9	51.6	50.5
BHMM	61.3	61.1	60.9	60.9	60.9
M^4	60.4	59.6	59.3	59.2	59.3
Twitter					
Unigram	93.0	93.0	93.0	93.0	93.0
LDA	83.7	78.4	74.0	70.9	70.2
BHMM	90.5	89.9	89.7	89.6	89.4
M^4	88.4	86.2	85.5	85.6	86.31

- M^4 more predictive than the block HMM