u+i lab
users and information

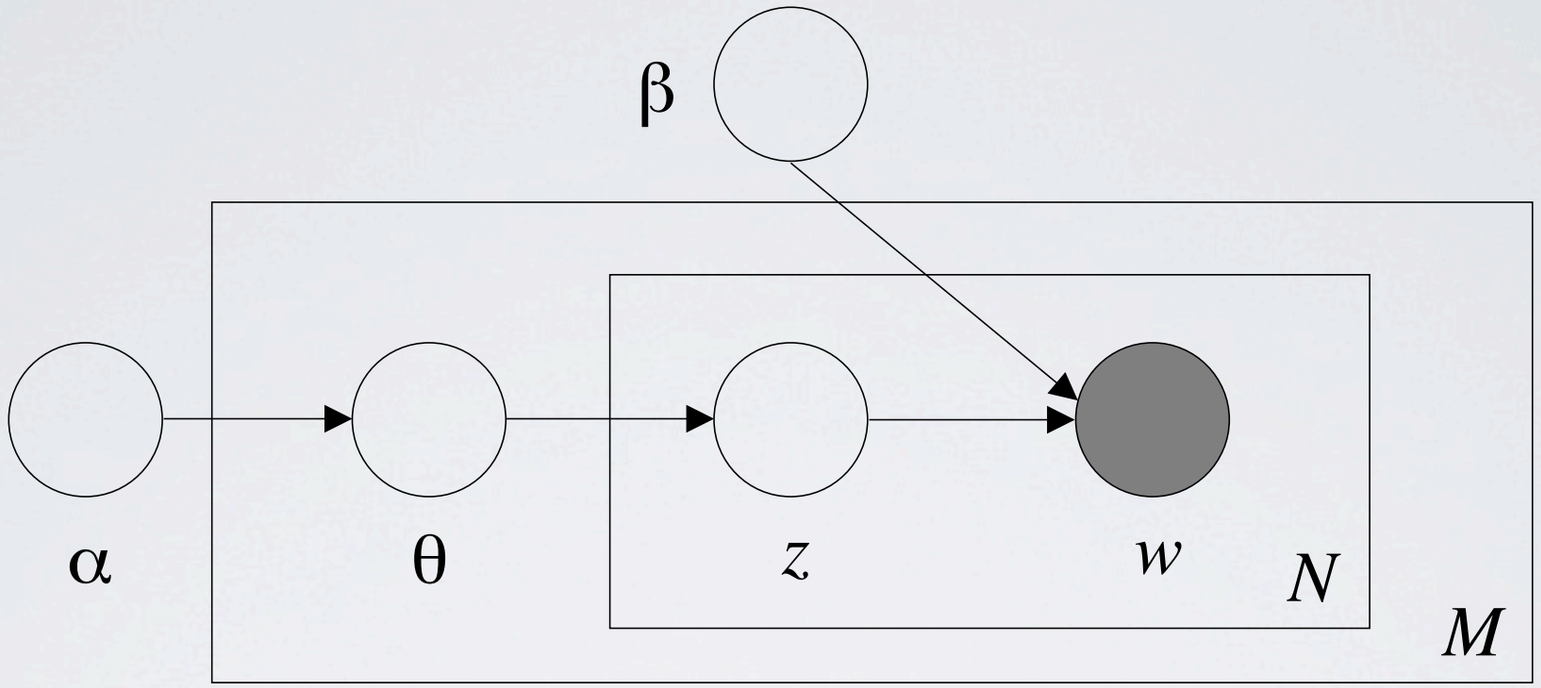
Distance-Dependent Chinese Restaurant Franchise

Dongwoo Kim
Computer Science Department

Thesis Presentation

Presentation Objectives

- Introduce the concept and problems of Bayesian *parametric / non-parametric* topic models
- Propose the distance-dependent Chinese restaurant franchise
- Present the experimental results
  - Based on four different time varying corpora
  - NIPS, SIGIR, SIGMOD, SIGGRAPH



Topic Model

An approach to analyze large volume of unlabeled documents

Concept of Topic Model

SIGIR corpus

Learning to Rank Only Using Training Data from Related Domain

Social Media Recommendation based on People and Tags

An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages

Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos and Constantine D. Spyropoulos
Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications
National Centre for Scientific Research "Demokritos"
153 10 Ag. Paraskevi, Athens, Greece
e-mail: {ionandr, jkoutsi, kostel, costass}@iit.demokritos.gr

Abstract

The growing problem of unsolicited bulk e-mail, also known as "spam", has generated a need for reliable anti-spam e-mail filters. Filters of this type have so far been based mostly on manually constructed keyword patterns. An alternative approach has recently been proposed, whereby a Naive Bayesian classifier is trained automatically to detect spam messages. We test this approach on a large collection of personal e-mail messages, which we make publicly available in "encrypted" form contributing towards standard benchmarks. We introduce appropriate cost-sensitive measures, investigating at the same time the effect of attribute-

Spam messages are annoying to most users, as they waste their time and clutter their mailboxes. They also cost money to users with dial-up connections, waste bandwidth, and may expose minors to unsuitable content (e.g. when advertising pornographic sites). A 1997 study [3] found that spam messages constituted approximately 10% of the incoming messages to a corporate network. The situation seems to be worsening, and without appropriate counter-measures, spam messages could eventually undermine the usability of e-mail.

Anti-spam legal measures are gradually being adopted, but they have had a very limited effect so far.[1] Of more direct value are *anti-spam filters*, software tools that attempt to block automatically spam messages.[2] Apart from blacklists of

Concept of Topic Model

An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages

Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos and Constantine D. Spyropoulos
Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications
National Centre for Scientific Research "Demokritos"
153 10 Ag. Paraskevi, Athens, Greece
e-mail: {ionandr, jkoutsi, kostel, costass}@iit.demokritos.gr

Abstract

The growing problem of unsolicited bulk e-mail, also known as "spam", has generated a need for reliable anti-spam e-mail filters. Filters of this type have so far been based mostly on manually constructed keyword patterns. An alternative approach has recently been proposed, whereby a Naive Bayesian classifier is trained automatically to detect spam messages. We test this approach on a large collection of personal e-mail messages, which we make publicly available in "encrypted" form contributing towards standard benchmarks. We introduce appropriate cost-sensitive measures, investigating at the same time the effect of attribute-set size, training-corpus size, lemmatization, and stop lists, issues that have not been explored in previous experiments.

Spam messages are annoying to most users, as they waste their time and clutter their mailboxes. They also cost money to users with dial-up connections, waste bandwidth, and may expose minors to unsuitable content (e.g. when advertising pornographic sites). A 1997 study [3] found that spam messages constituted approximately 10% of the incoming messages to a corporate network. The situation seems to be worsening, and without appropriate counter-measures, spam messages could eventually undermine the usability of e-mail.

Anti-spam legal measures are gradually being adopted, but they have had a very limited effect so far.[1] Of more direct value are *anti-spam filters*, software tools that attempt to block automatically spam messages.[2] Apart from blacklists of frequent spammers and lists of trusted users, which can be

- Every word has its latent topic (unknown)

spam filtering

experiment

ML

Concept of Topic Model

An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages

Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos and Constantine D. Spyropoulos
Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications
National Centre for Scientific Research "Demokritos"
153 10 Ag. Paraskevi, Athens, Greece
e-mail: {ionandr, jkoutsi, kostel, costass}@iit.demokritos.gr

Abstract

The growing problem of unsolicited bulk e-mail, also known as "spam", has generated a need for reliable anti-spam e-mail filters. Filters of this type have so far been based mostly on manually constructed keyword patterns. An alternative approach has recently been proposed, whereby a Naive Bayesian classifier is trained automatically to detect spam messages. We test this approach on a large collection of personal e-mail messages, which we make publicly available in "encrypted" form contributing towards standard benchmarks. We introduce appropriate cost-sensitive measures, investigating at the same time the effect of attribute-set size, training-corpus size, lemmatization, and stop lists, issues that have not been explored in previous experiments.

Spam messages are annoying to most users, as they waste their time and clutter their mailboxes. They also cost money to users with dial-up connections, waste bandwidth, and may expose minors to unsuitable content (e.g. when advertising pornographic sites). A 1997 study [3] found that spam messages constituted approximately 10% of the incoming messages to a corporate network. The situation seems to be worsening, and without appropriate counter-measures, spam messages could eventually undermine the usability of e-mail.

Anti-spam legal measures are gradually being adopted, but they have had a very limited effect so far.[1] Of more direct value are *anti-spam filters*, software tools that attempt to block automatically spam messages.[2] Apart from blacklists of frequent spammers and lists of trusted users, which can be

- Topic modeling task can be viewed as a **topic assignment to every word**

Output of Topic Model

spam	0.12		machine	0.10	
email	0.10		learning	0.09	
filtering	0.08		model	0.08	
filters	0.06		likelihood	0.08	
filter	0.06		class	0.07	
messages	0.05		variable	0.06	
:	:		:	:	
Bayesian	0.001		spam	0.001	

experiment	0.15
validation	0.11
result	0.9
performance	0.8
test	0.7
perform	0.6
:	:
email	0.001

Topics: multinomial over vocabulary

Parametric Topic Models

- Representative model is LDA(Latent Dirichlet Allocation)(Blei, 2004)
- Number of topics must be **determined** for the corpus
  - Like K-means, user should select appropriate number of topics K before training the model

A Limitation of Parametric Topic Models

- It is difficult to determine the number of topics for a corpus
  - Computing optimal number of topics is very time-consuming
  - The optimal number of topics varies for each corpus

Non-Parametric Topic Models

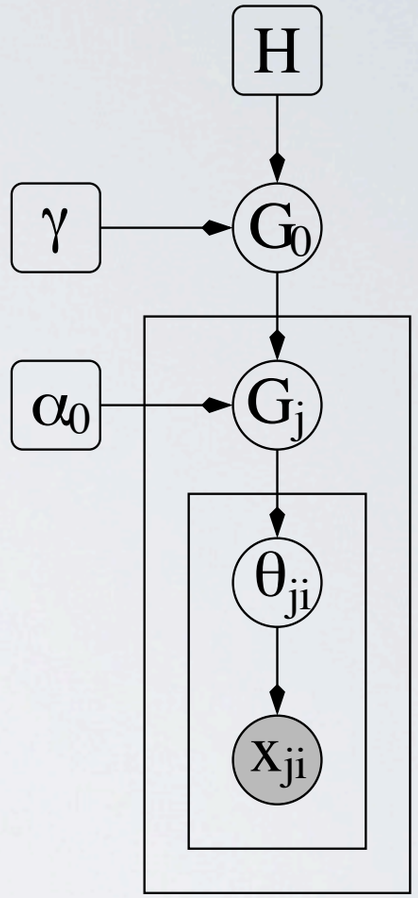
- Representative model is HDP-LDA(Hierarchical Dirichlet Process)(Teh, 2006)
- Assumes an **infinite** number of topics
- Model automatically captures the appropriate number of topics

A Limitation of HDP

- HDP does not consider the relationships among the documents
- However, some document collections exhibit patterns arising from the relationships among the documents
- For example, articles from conference proceedings exhibit a temporal pattern of topics
- When assigning topics to documents, probabilities for topics within a nearby neighborhood of documents should be higher than the topics in documents that are far apart

Proposed Model

- This thesis proposes a variant of HDP, called distance-dependent Chinese restaurant franchise
- ddCRF considers the relationships among the documents in the corpus
- ddCRF captures the temporal patterns of topics within conference proceedings



Distance-Dependent Chinese Restaurant Franchise

Variation of Bayesian non-parametric topic model

u+i lab
users and information

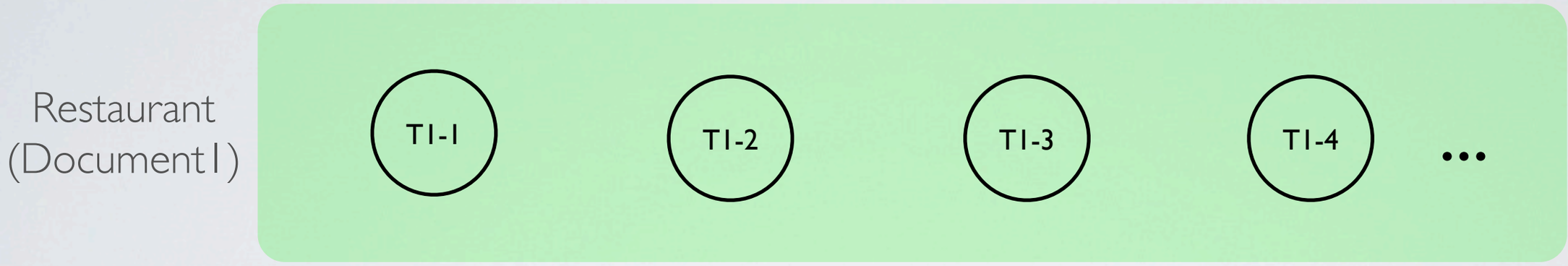
Chinese Restaurant Franchise (CRF)

- HDP is hard to imagine & understand
- Introduced as a Metaphor to explain the HDP(Teh, 2006)
- Two level hierarchical Chinese restaurant process(CRP)

CRF Metaphor

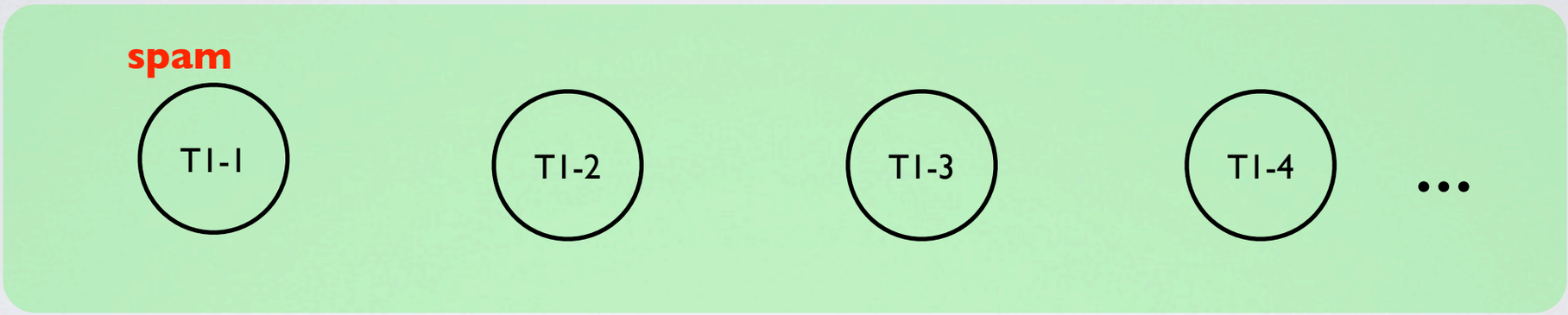
- Each restaurant has an infinite number of tables
- *N* customers are sequentially sitting down at the tables
- Each table has one dish to serve

CRF to Topic Model



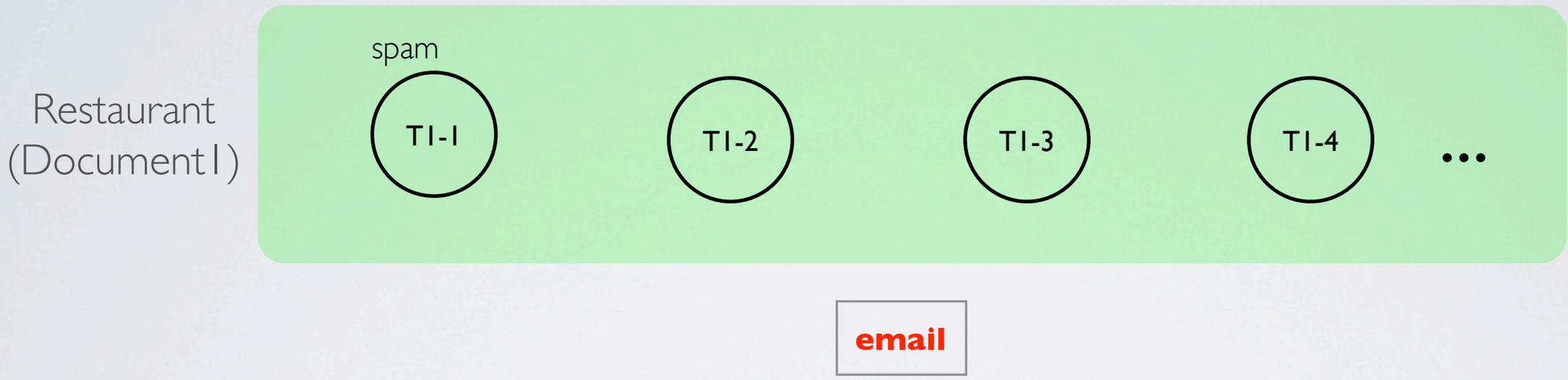
- Explain topic modeling by using CRF metaphor
  - Consider a restaurant as a document
  - Consider a customer as a word
  - Consider a table as a topic

CRF to Topic Model



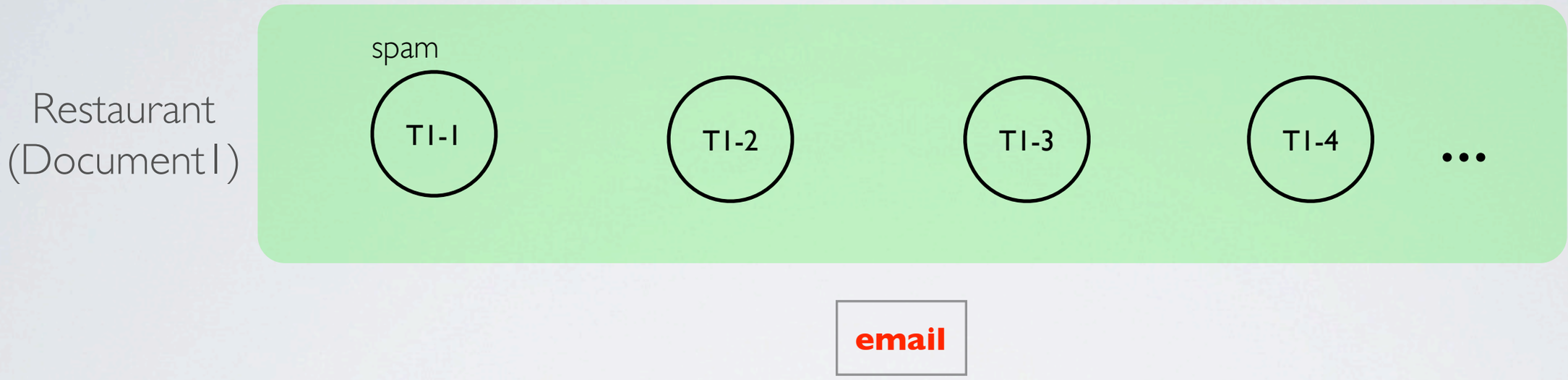
First customer '*spam*' is coming to '*document1*' restaurant
And sitting at the table T1-1

CRF to Topic Model



Second customer '*email*' is coming to '*document 1*' restaurant
And considering where to sit

CRF to Topic Model



- Probability of *'email'* sitting at

  ➡ an occupied table is proportional to the number of customers already sitting at that table

  ➡ new table is proportional to a constant γ

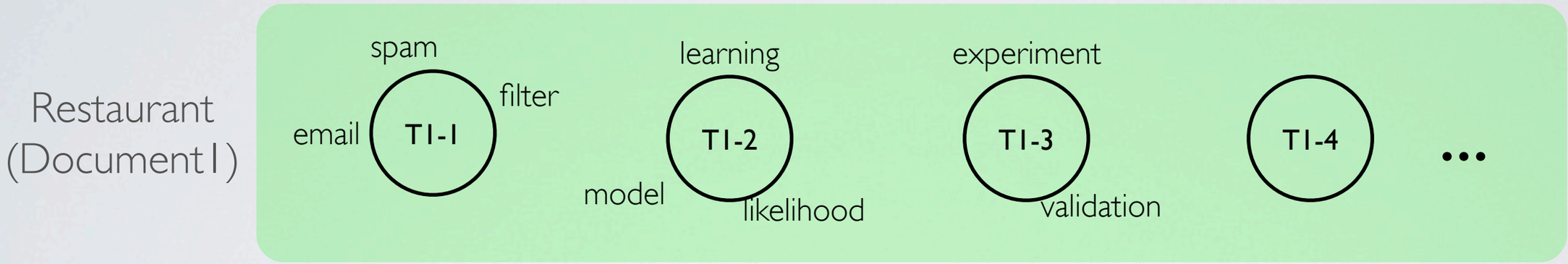
CRF to Topic Model

email

- Formally, probability of *'email'* sitting at the

occupied table *k*	$p(z_i = k \mid z_{1:(i-1)}, \gamma) = \frac{n_k}{\gamma + i - 1}$	$z_i =$ table no of ith customer
new table	$p(z_i = K+1 \mid z_{1:(i-1)}, \gamma) = \frac{\gamma}{\gamma + i - 1}$	$n_k =$ number of customers already sitting at table k
		$\gamma =$ parameter

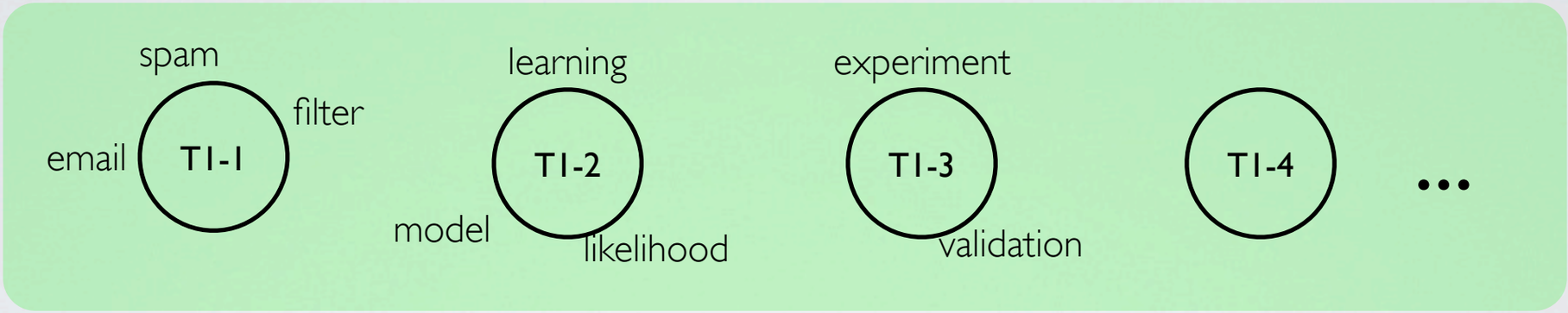
CRF to Topic Model



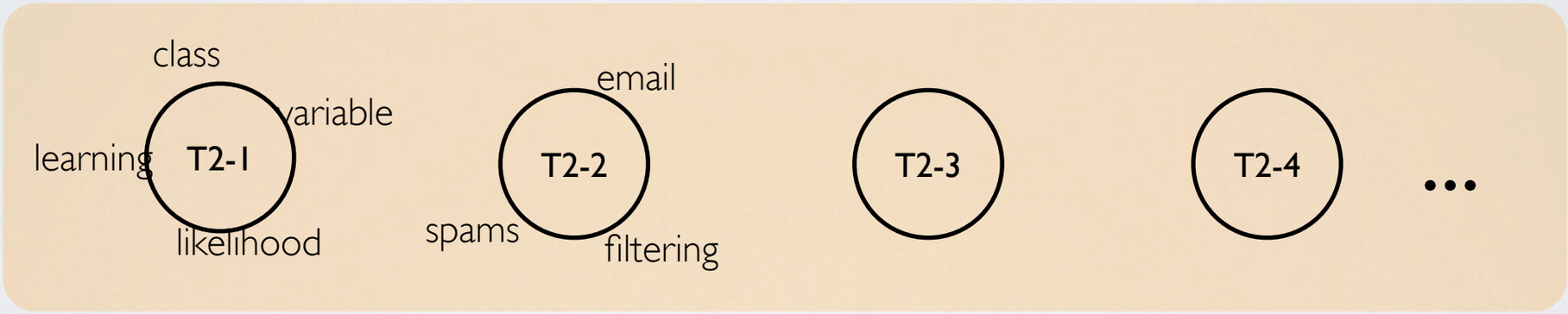
- Above result shows
  - configuration after *N* customers are sitting at the tables
- This process represents how CRP works

CRF to Topic Model

Restaurant (Document1)



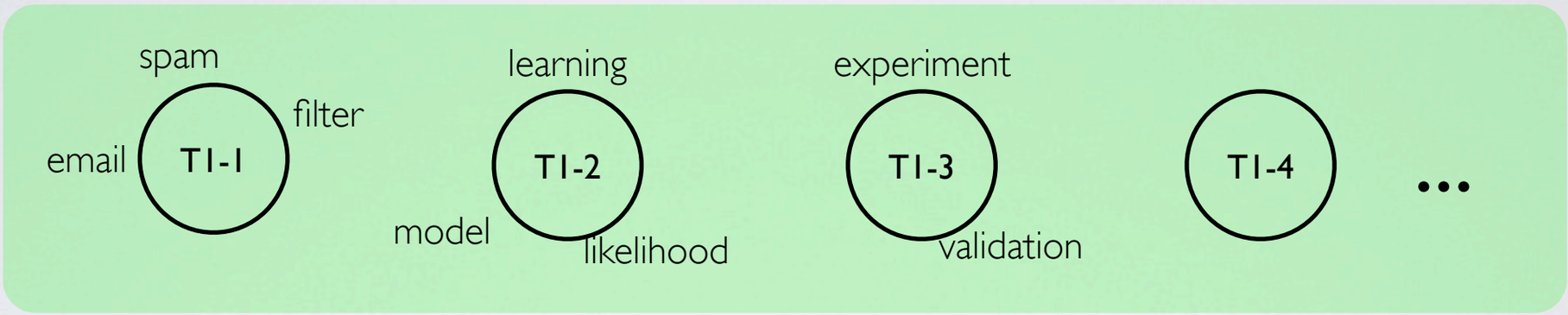
Restaurant (Document2)



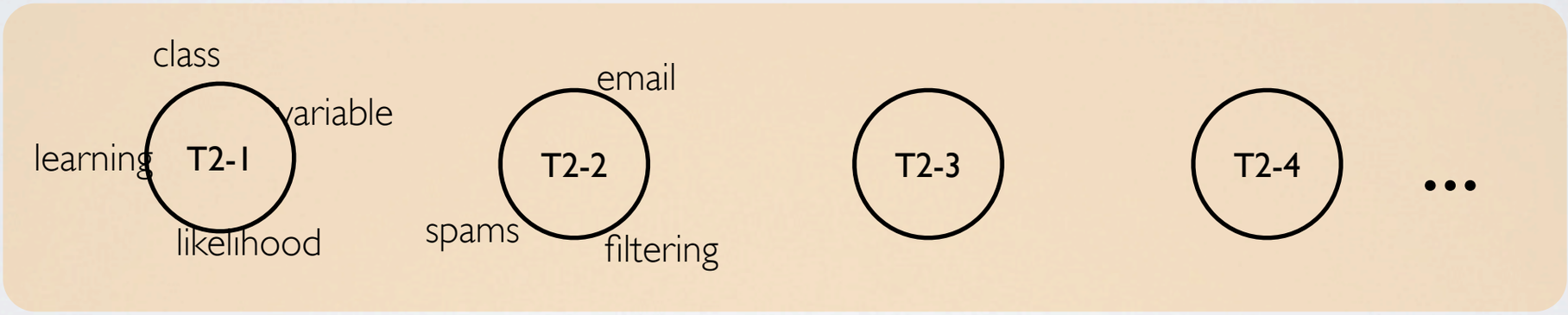
There are many restaurants in the world !!!

CRF to Topic Model

Restaurant (Document1)



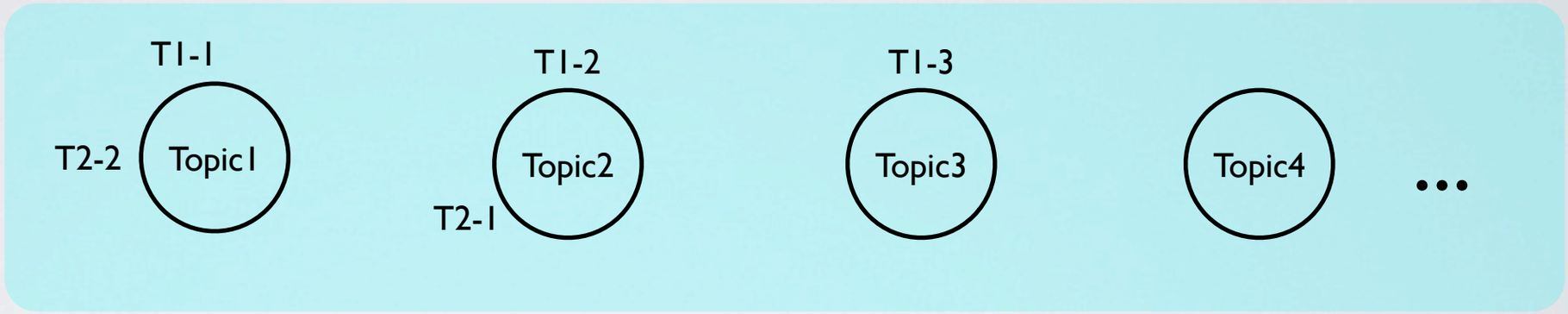
Restaurant (Document2)



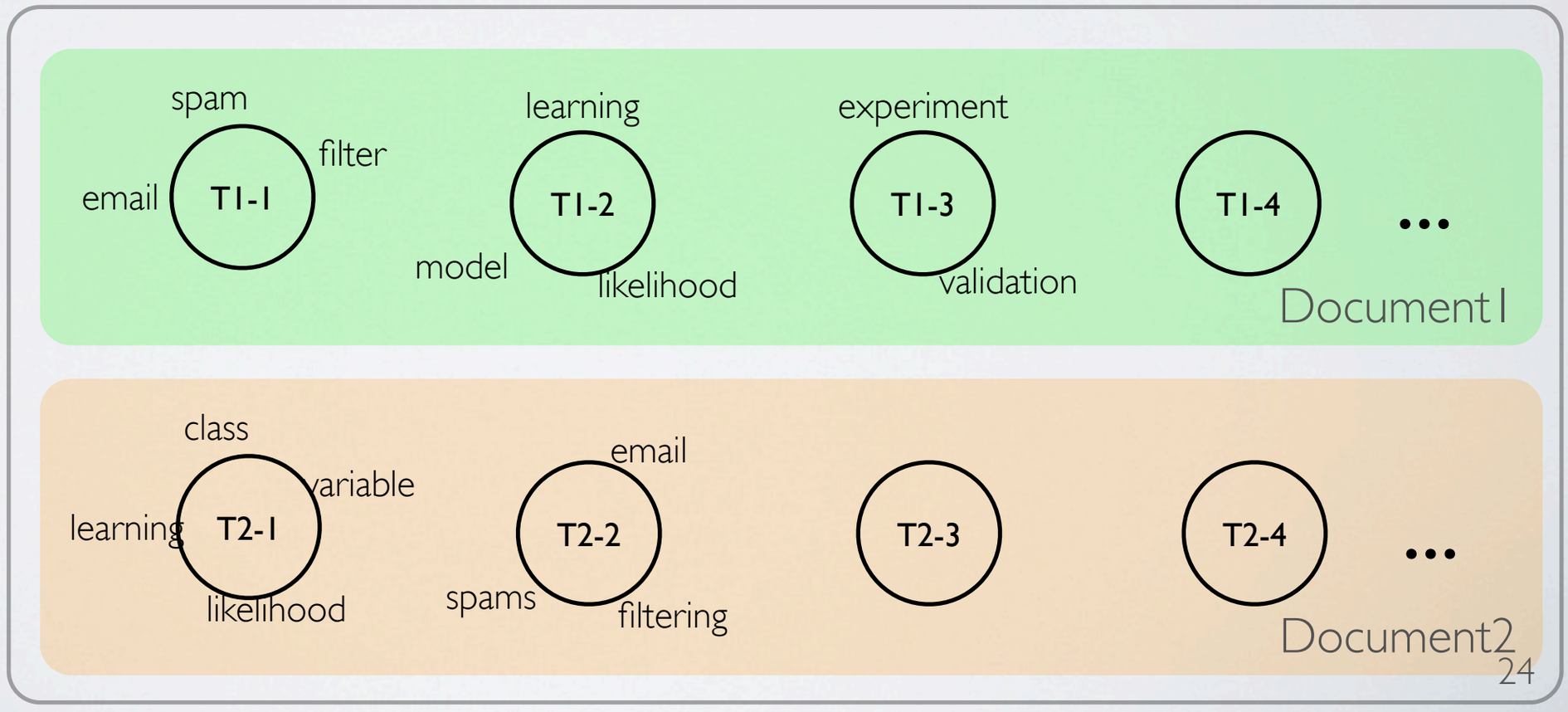
However, how do we know that T1-1 and T2-2 are the same topics?

Introduce Menu Level CRP

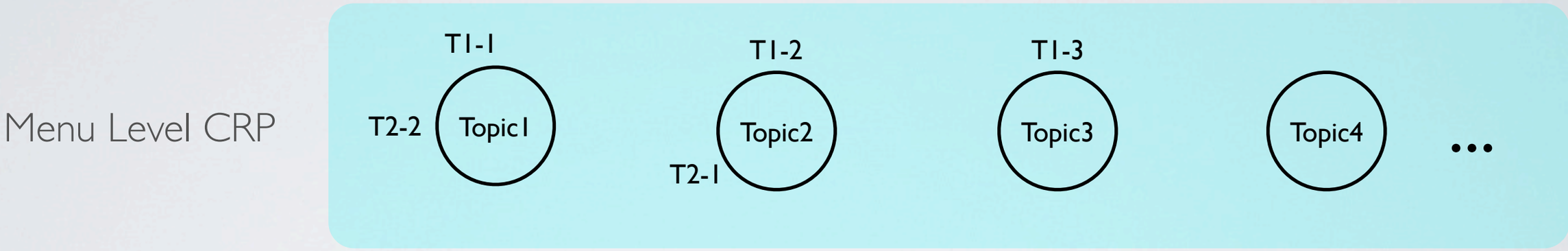
Menu Level CRP



Customer Level CRP

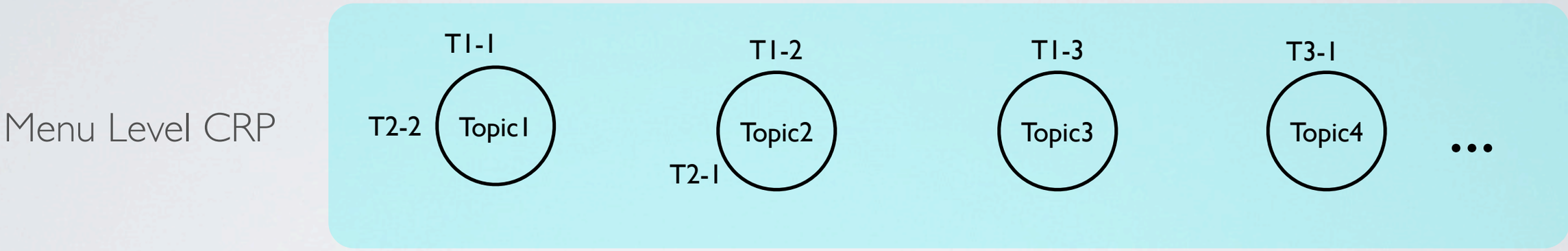


Introduce Menu Level CRP



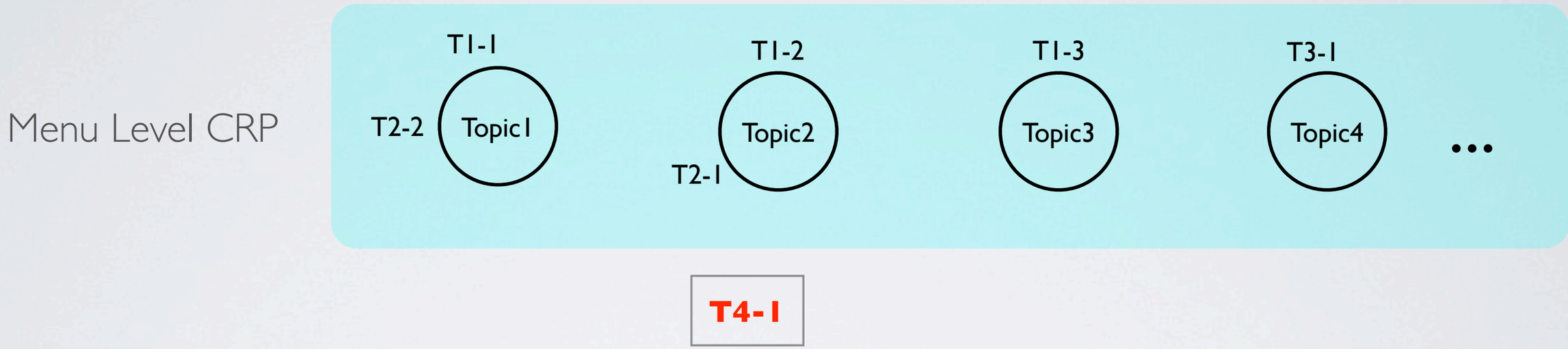
- Menu level CRP decides which dish is served to each table
  - Customers at Table 'T1-1' and 'T2-2' eat dish 'Topic1'
  - Customers at Table 'T1-2' and 'T2-1' eat dish 'Topic2'

Limitation of Menu Level CRP



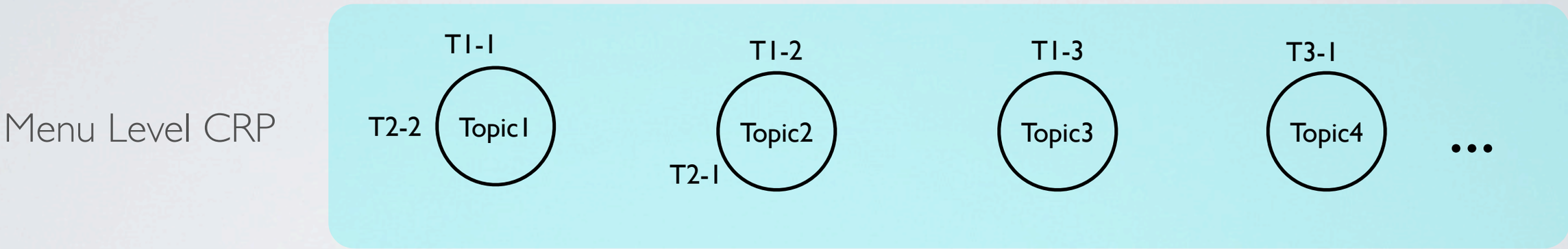
- New Assumptions
  - Document 1 and document 2 are written in 2000
  - Document 3 is written in 1978

Limitation of Menu Level CRP



- Topic2 is about '*spam filtering*'
- Now, new table 'T4-1' is coming and choosing a menu to serve
  - What if table 'T4-1' is a document written in 1979

Limitation of Menu Level CRP

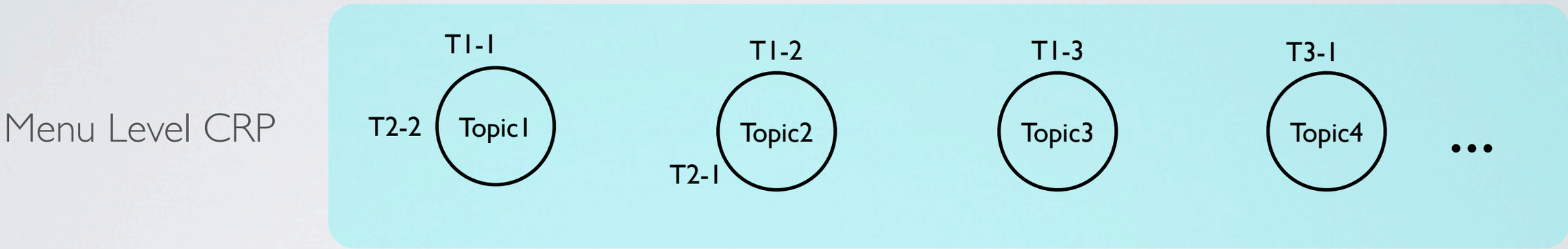


- Original CRF does not consider a relationship between tables
  - Topic2(*spam-filtering*) can be served to the table T4-1(*1979 document*)
  - This is not an appropriate modeling

Introduce ddCRF Metaphor

- Selection of a dish could be influenced by nearby restaurants
- If there is a famous menu in a specific region, we probably want to eat that menu in that region
- If a document was written in 1979, the topics would be more likely to be the same as the documents written in 1978, and less likely to be the same as the documents written in 2000

Consider Relationship



T4-1

- For choosing dish for table 'T4-1', we compare the relationship between 'T4-1' and others tables already sitting down at menu level tables

Distance-Dependent CRP

$$p(z_i = k|\ z_{1:(i-1)}, \gamma) = \frac{n_k}{\gamma + i - 1}$$

$$p(z_i = K+1|\ z_{1:(i-1)}, \gamma) = \frac{\gamma}{\gamma + i - 1}$$

$$p(z_i = k \mid D, z_{1:(i-1)}, \gamma) = \frac{\sum_{z_j = k} f(d_{ij})}{\gamma + \sum_{j \neq i} f(d_{ij})}$$

$$p(z_i = K+1 \mid D, z_{1:(i-1)}, \gamma) = \frac{\gamma}{\gamma + \sum_{j \neq i} f(d_{ij})}.$$

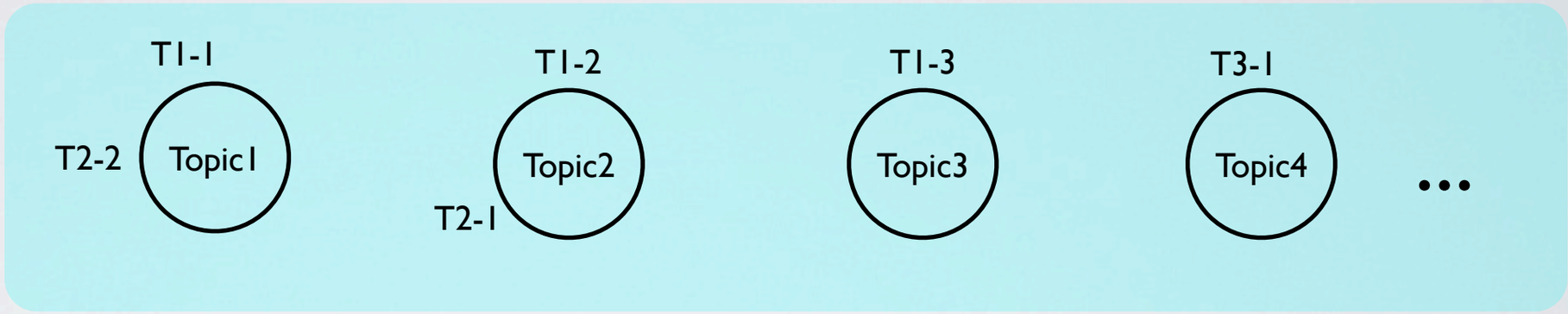
$f = \text{decay function}$

$d_{ij} = \text{distance between } i \text{ \& } j$

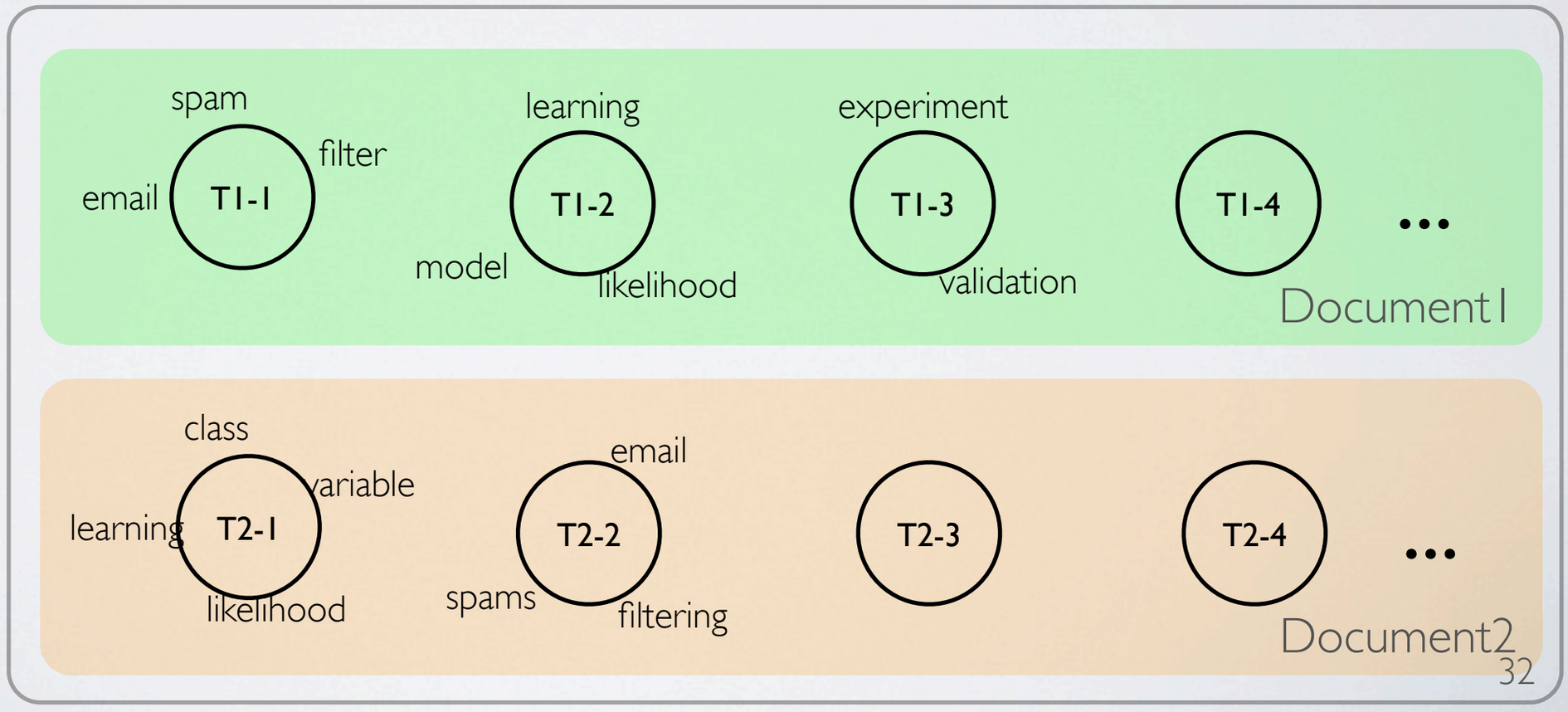
- We model the relationship as a distance between documents
- Distance metric should be defined such that the distance between two close documents is small
- Decay function
  - 0(long distance) ~ 1(short distance)

Distance-Dependent CRF

Menu Level
ddCRP



Customer Level
CRP



u+i lab
users and information

Experiments

Comparisons with other topic models

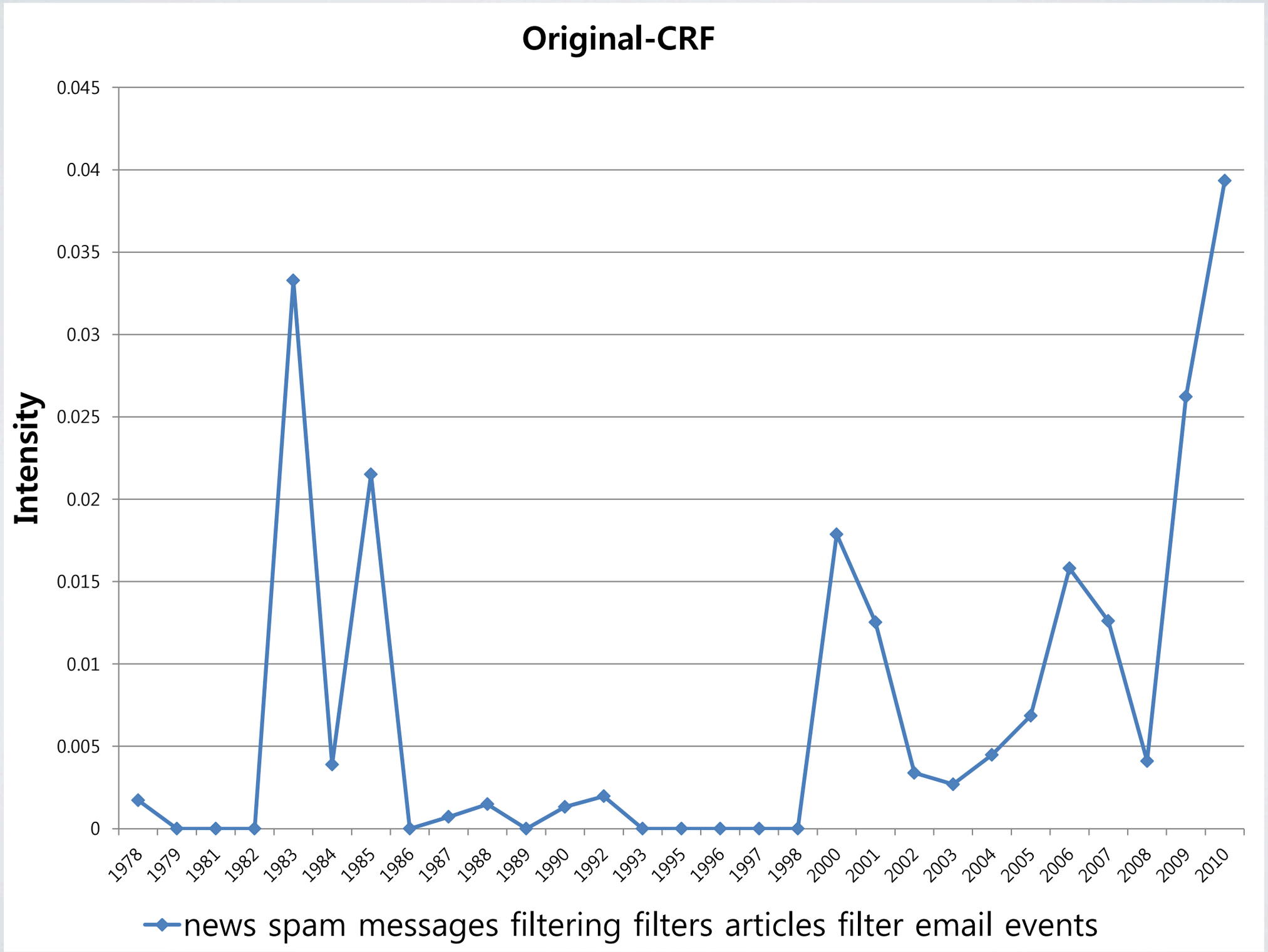
Experiments

- Infer topics from four different corpora (SIGIR, SIGMOD, SIGGRAPH, NIPS)
  - Use MCMC (Gibbs-Sampling) technique for inference
- Use 2 kinds of decay functions
  - Logistic decay function, exponential decay function
- Compared with other topic models
  - LDA, HDP

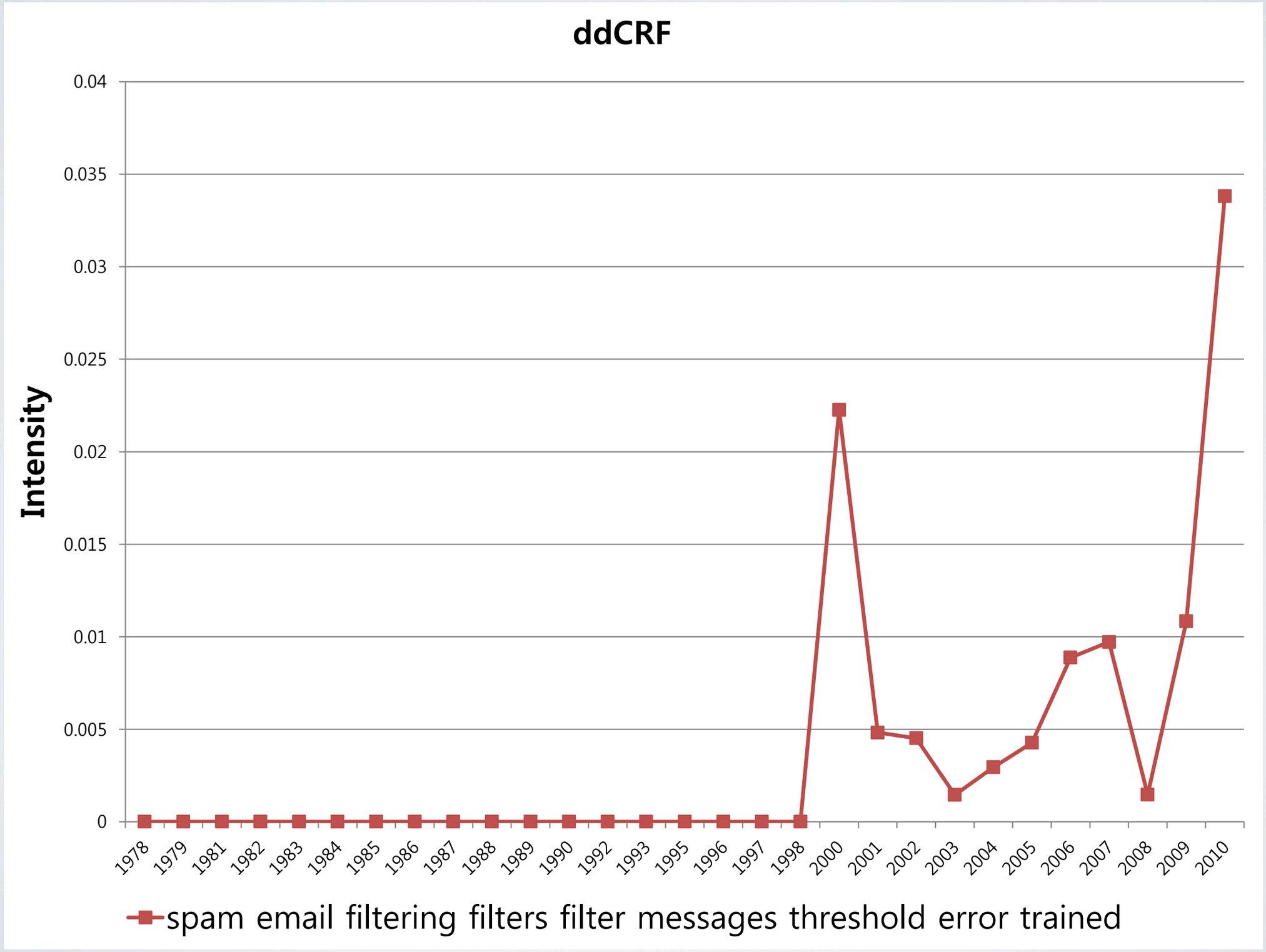
Evaluation Metrics

- Qualitative Evaluation
  - Emergence and disappearance of topics
- Quantitative Evaluation
  - Held-out likelihood
  - Complexity

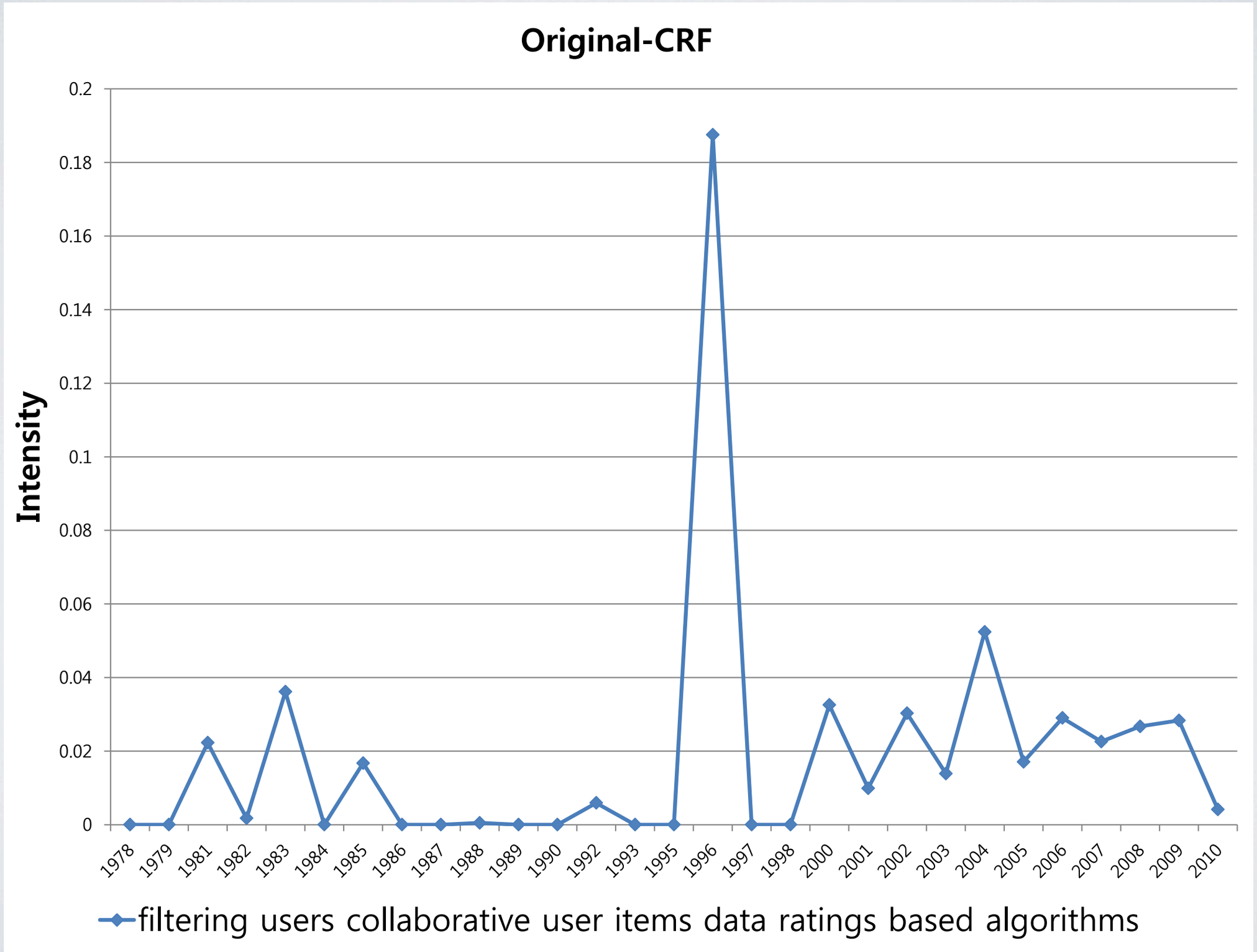
Topic Emergence



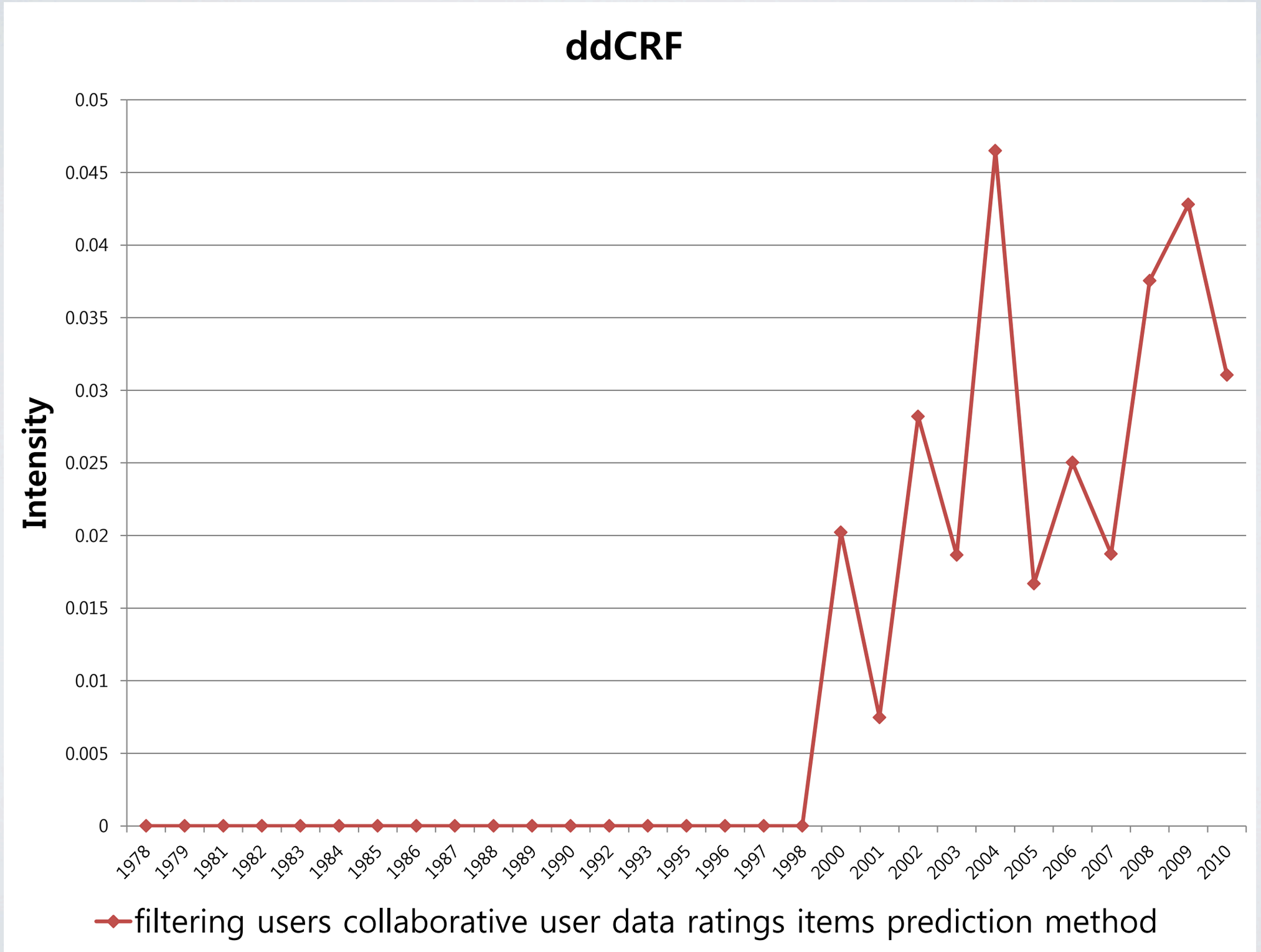
Topic Emergence



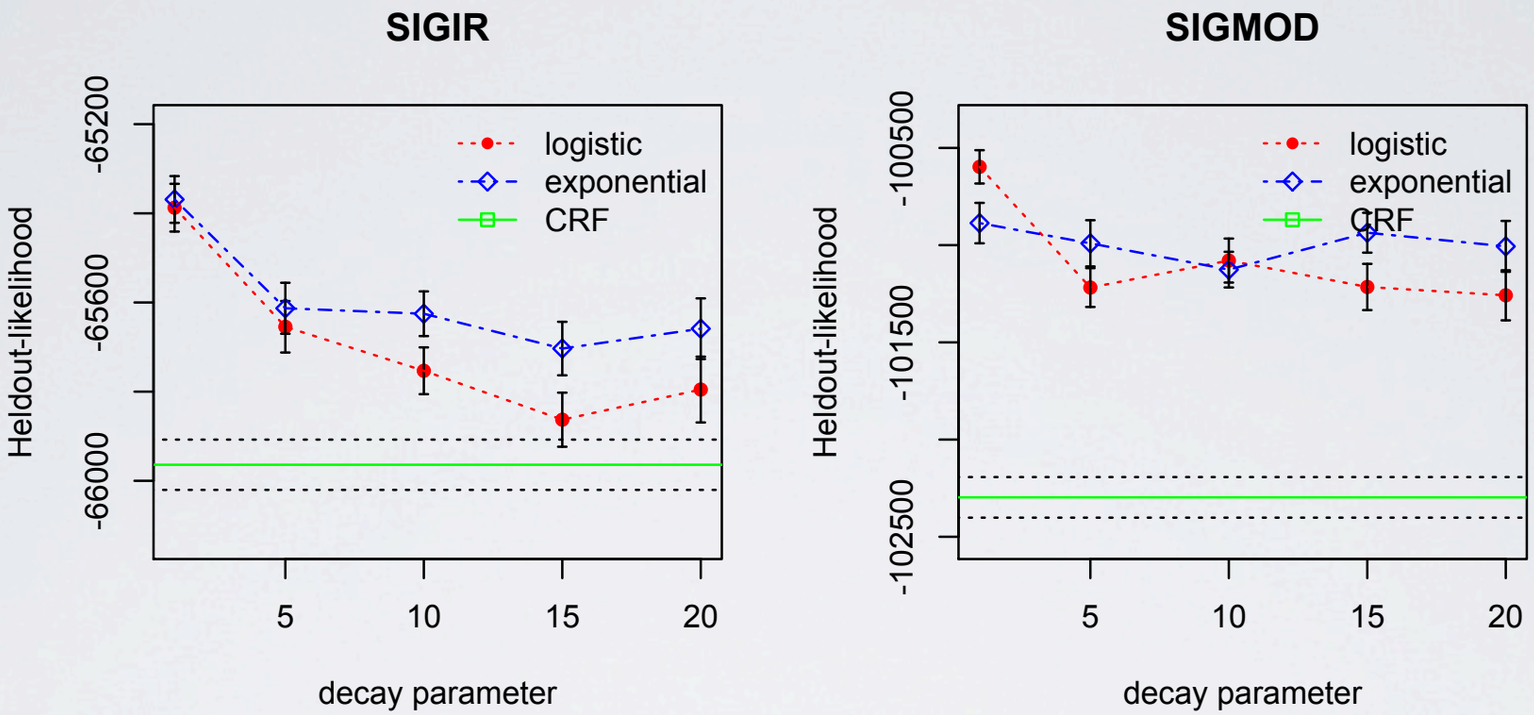
Topic Emergence

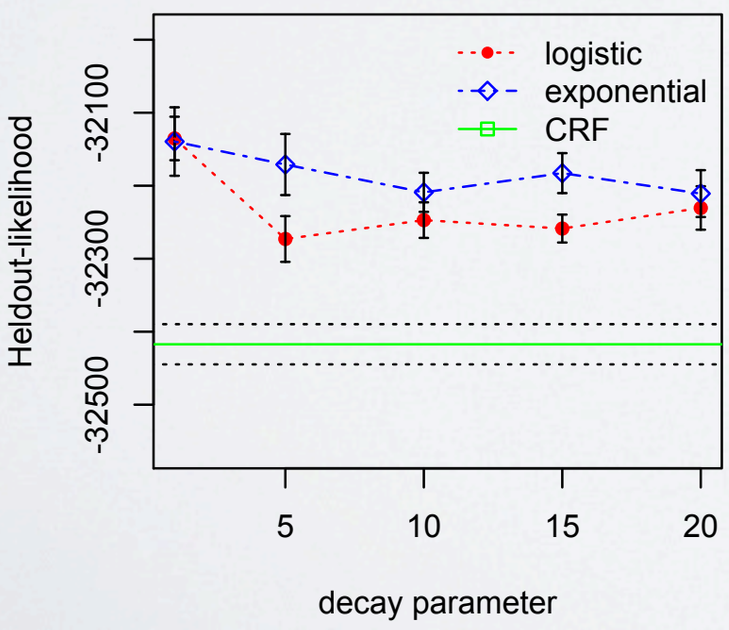


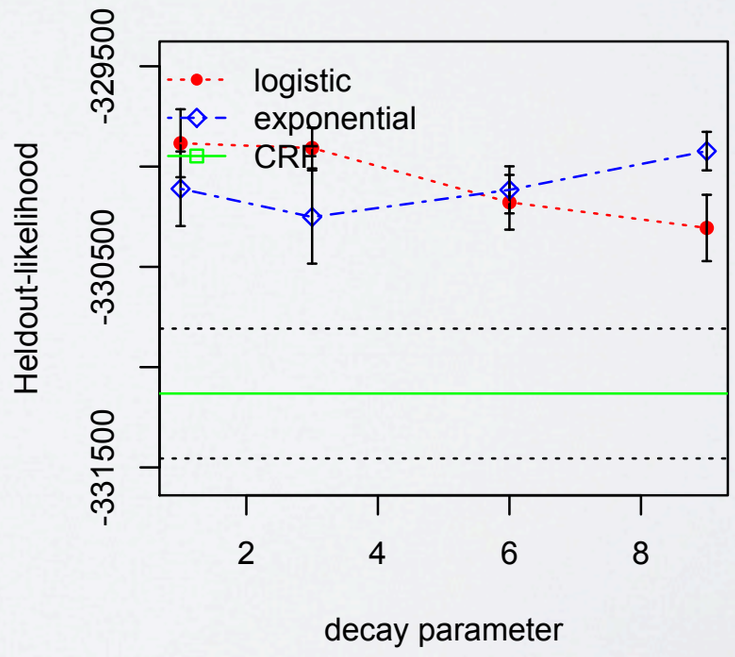
Topic Emergence



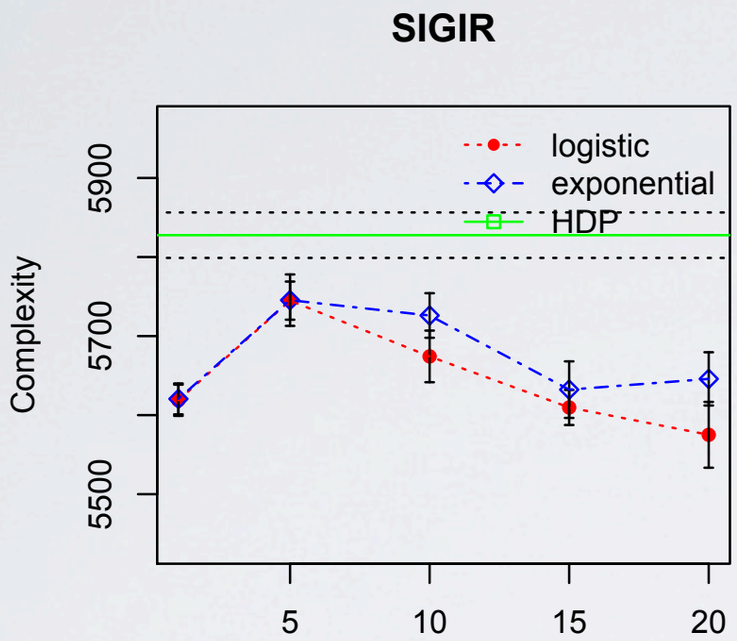
Held-out Likelihood

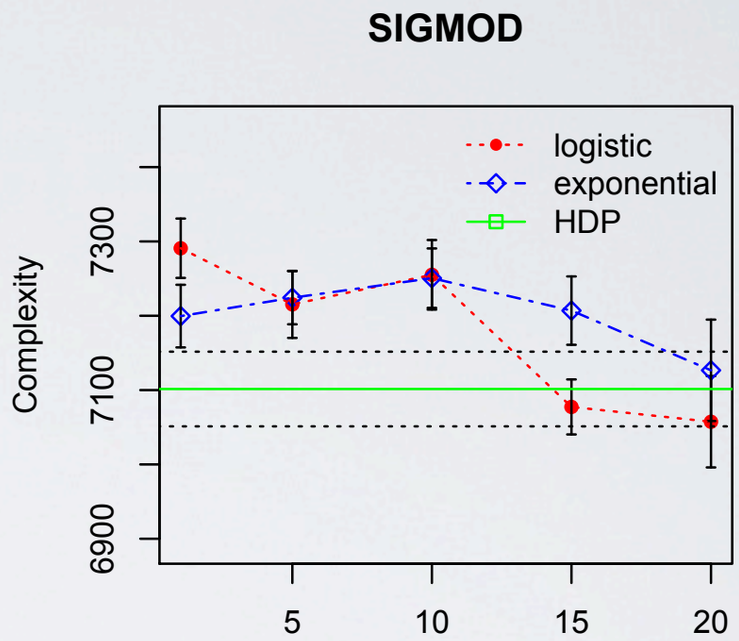


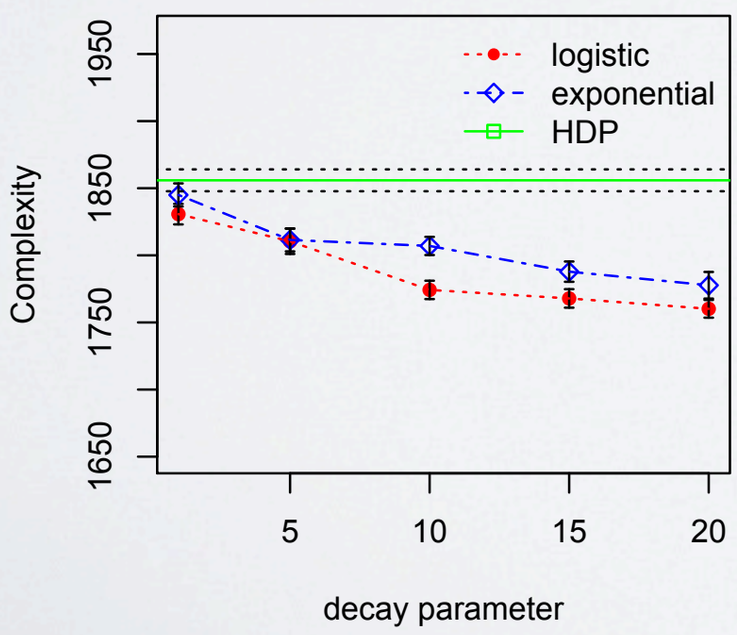


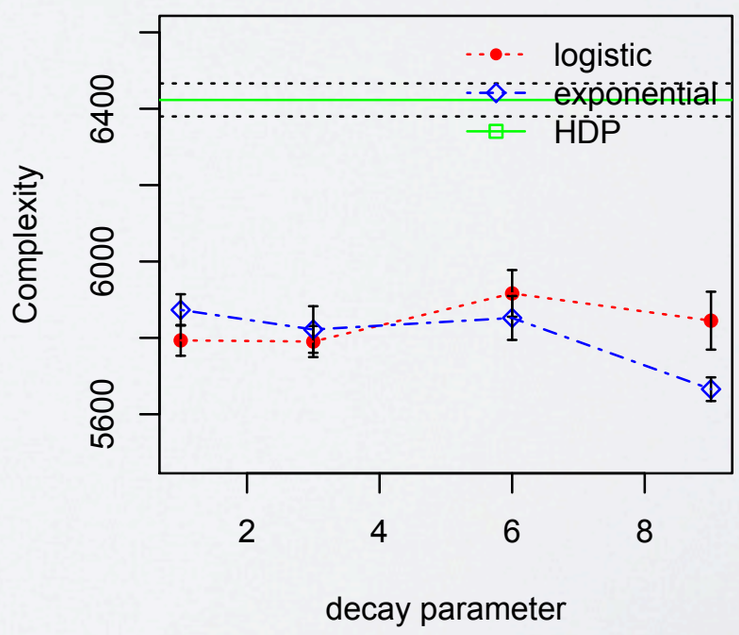


Complexity


SIGIR
Complexity
5500
5700
5900
5
10
15
20
decay parameter
logistic
exponential
HDP


SIGMOD
Complexity
6900
7100
7300
5
10
15
20
decay parameter
logistic
exponential
HDP


SIGGRAPH
Complexity
1650
1750
1850
1950
5
10
15
20
decay parameter
logistic
exponential
HDP


NIPS
Complexity
5600
6000
6400
2
4
6
8
decay parameter
logistic
exponential
HDP

Contributions & Future Work

- Designed and implemented distance-dependent CRF
  - A variant of Chinese restaurant franchise
  - For a corpus where the relationships among documents are important
- Modeled topics from four different corpora to capture temporal patterns of topics
  - Quantitative evaluation shows improved performance over LDA(parametric topic model) and HDP(non-parametric topic model)
  - Qualitative evaluation shows interesting temporal patterns of topic emergence
- Future work will explore various definitions of distance: time dimension, spatial dimension, or some other dimension
- The ddCRF can be applied to various other problems where other topic models have been successfully applied
  - Cognitive science, computational biology, multimedia (image, music, video) analysis ...

Q&A