YAHOO!®

Storylines from Streaming Text

The Infinite Topic Cluster Model

Amr Ahmed, Jake Eisenstein, Qirong Ho
Alex Smola, Choon Hui Teo, Eric Xing

Carnegie Mellon University Yahoo! Research

Outline

- Visualizing a news stream
- Goals

  Clusters & Content analysis
- The story-topic model

  Recurrent Chinese Restaurant Process

  Latent Dirichlet Allocation
- Examples

News Stream

AP

Republicans is bec
and lawmakers. Bu
Bill Clinton even ba
Full Story »

Video: Gibbs: I Hav
Slideshow: Preside
Related: Tax fight

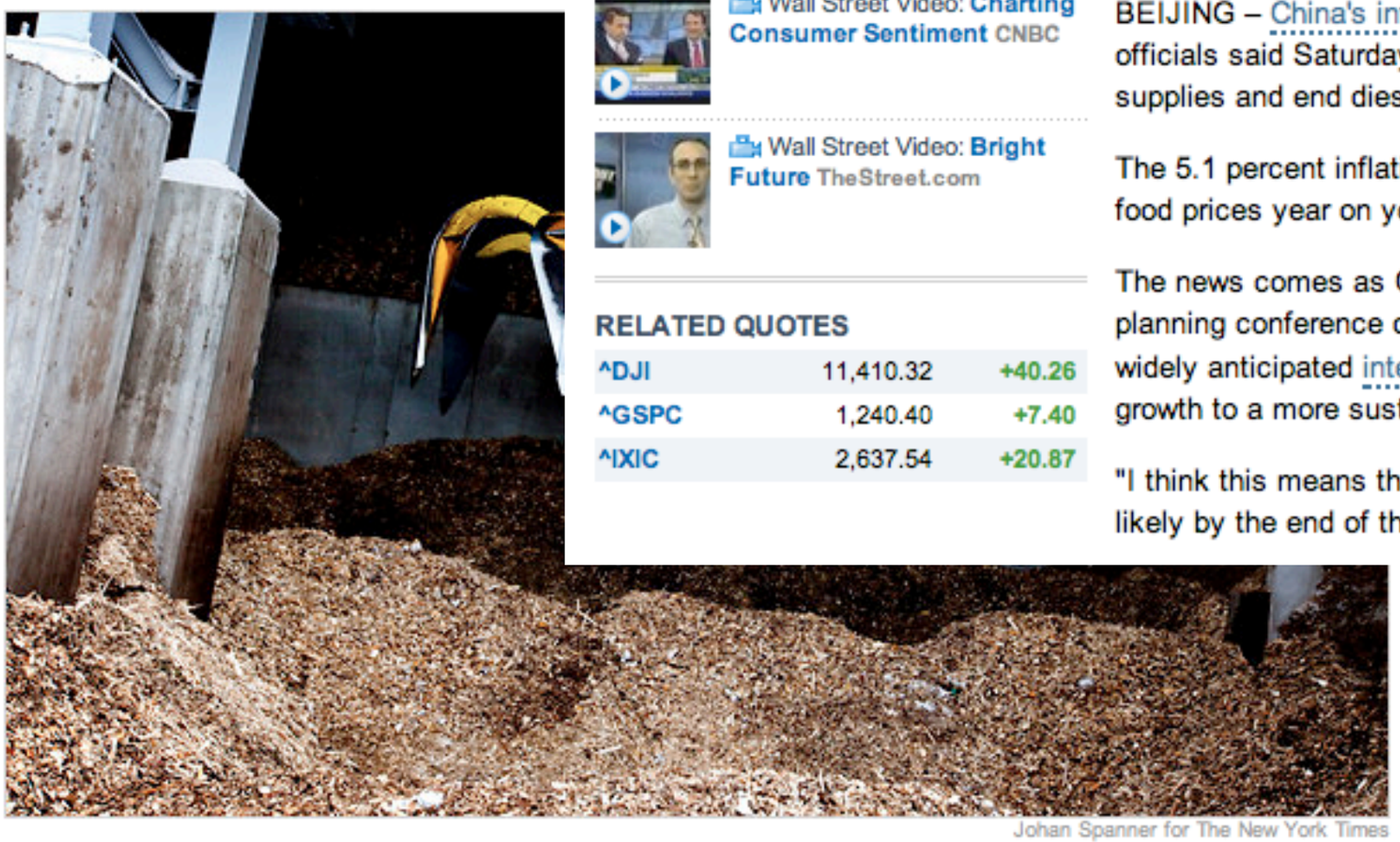
Add-ons turn tax cut bill into 'Christmas t

AP – 1 hr 32 mins ago

WASHINGTON – In the

BEYOND FOSSIL FUELS

Using Waste, Swedish C



Johan Spanner for The New York Times

As part of its citywide system, Kristianstad burns wood waste like tree prunings and scraps from flooring factories to power an underground district heating grid.

China says inflation up 5.1 perce

Associated Press

Buzz up! 19 votes | Share

By CARA ANNA, Associated Press

BEIJING – China's inflation su
officials said Saturday, despit
supplies and end diesel shorta

Wall Street Video: **Charting Consumer Sentiment** CNBC

Wall Street Video: **Bright Future** TheStreet.com

RELATED QUOTES

^DJI	11,410.32	+40.26
^GSPC	1,240.40	+7.40
^IXIC	2,637.54	+20.87

The 5.1 percent inflation rate was driven by a 11.7 percent jump in food prices year on year.

The news comes as China's leaders meet for the top economic planning conference of the year and as financial markets watch for a widely anticipated interest rate hike to help bring rapid economic growth to a more sustainable level.

"I think this means that an interest rate hike of 25 basis points is very likely by the end of the year," said CLSA analyst Andy Rothman.

Suit to Recover Madoff's Money Calls Austrian an Accomplice

By DIANA B. HENRIQUES and PETER LATTMAN

Sonja Kohn, an Austrian banker, is accused of masterminding a 23-year conspiracy that played a central role in financing the gigantic Ponzi scheme.

Post a Comment

Print

ber

ovember,

ase food

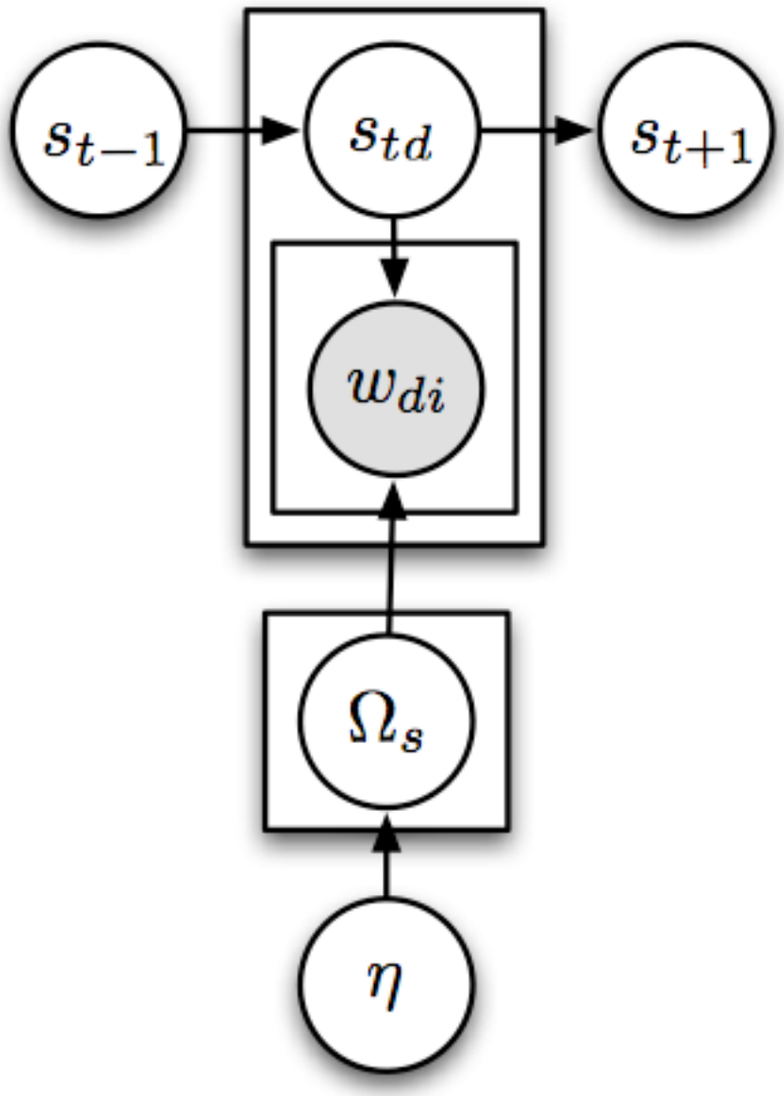
News Stream

- Realtime news stream
  - Multiple sources (Reuters, AP, CNN, …)
  - Same story from multiple sources
  - Stories are related

- Goals
  - Aggregate articles into a storyline
  - Analyze the storyline (topics, entities)

Precursors

Evolutionary Clustering / RCRP



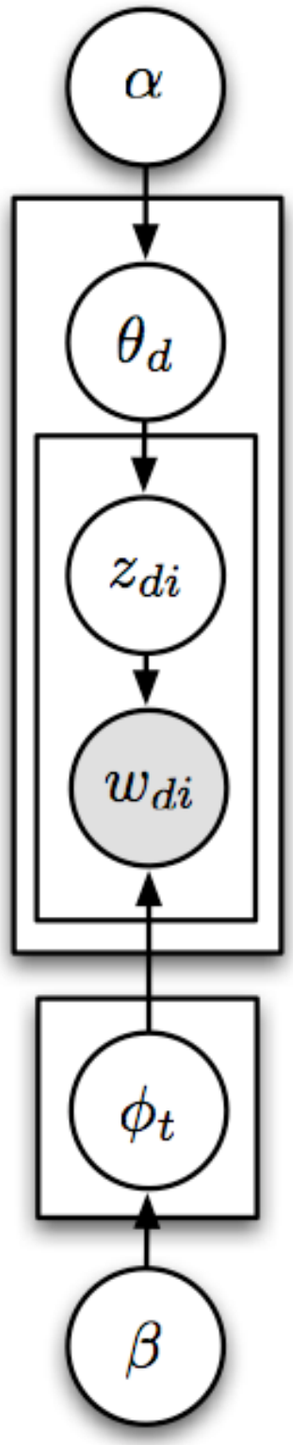
- Assume active story distribution at time t
- Draw story indicator
- Draw words from story distribution
- Down-weight story counts for next day

Ahmed & Xing, 2008

Clustering / RCRP

- Pro
  - Nonparametric model of story generation (no need to model frequency of stories)
  - No fixed number of stories
  - Efficient inference via collapsed sampler
- Con
  - We learn nothing!
  - No content analysis

Latent Dirichlet Allocation



- Generate topic distribution per article
- Draw topics per word from topic distribution
- Draw words from topic specific word distribution

Blei, Ng, Jordan, 2003

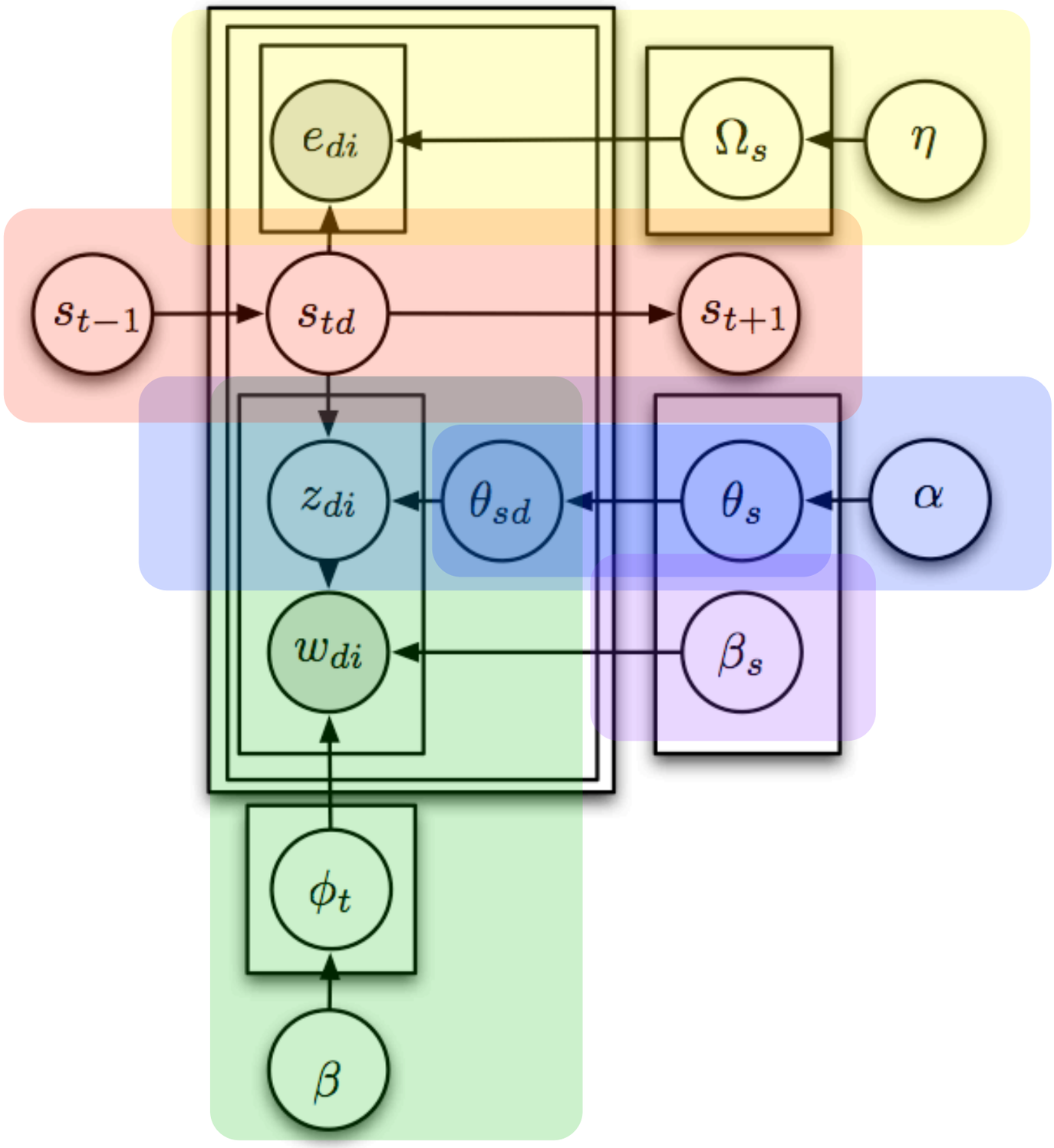
Latent Dirichlet Allocation

- Pro
  - Topical analysis of stories
  - Topical analysis of words (meaning, saliency)
  - More documents improve estimates
- Con
  - No clustering

More Issues

Storylines

Storylines Model



- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

Inference

- We receive articles as a stream
  Want topics & stories now
- Variational inference infeasible
  (RCRP, sparse to dense, vocabulary size)
- We have a ‘tracking problem’
  - Sequential Monte Carlo
  - Use sampled variables of surviving particle
  - Use ideas from Cannini et al. 2009

Estimation

- Proposal distribution - draw stories s, topics z

$$p(s_{t+1}, z_{t+1} | x_{1\ldots t+1}, s_{1\ldots t}, z_{1\ldots t})$$

using Gibbs Sampling for each particle

- Reweight particle via

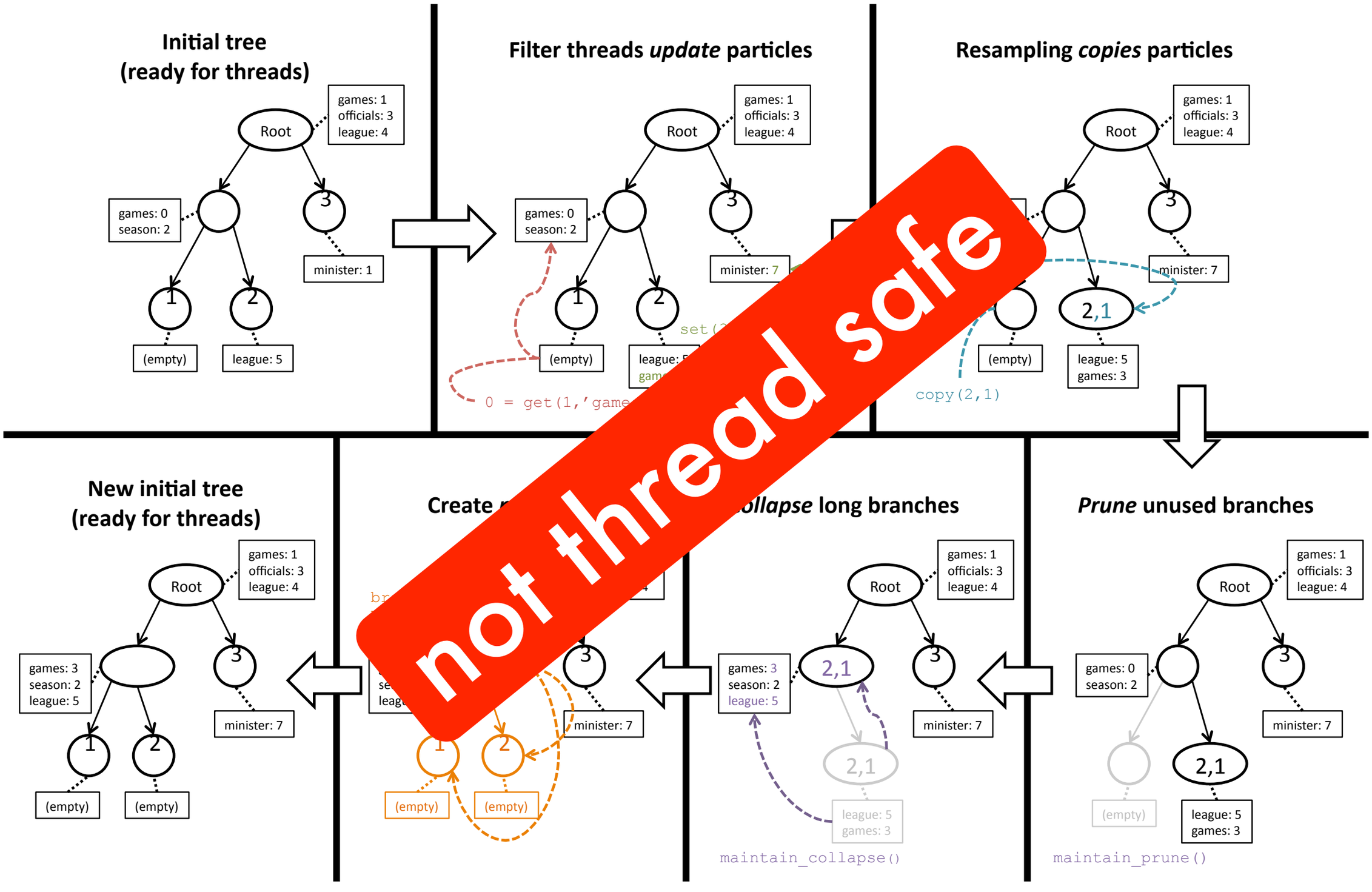
$$p(x_{t+1} | x_{1\ldots t}, s_{1\ldots t}, z_{1\ldots t})$$

new data

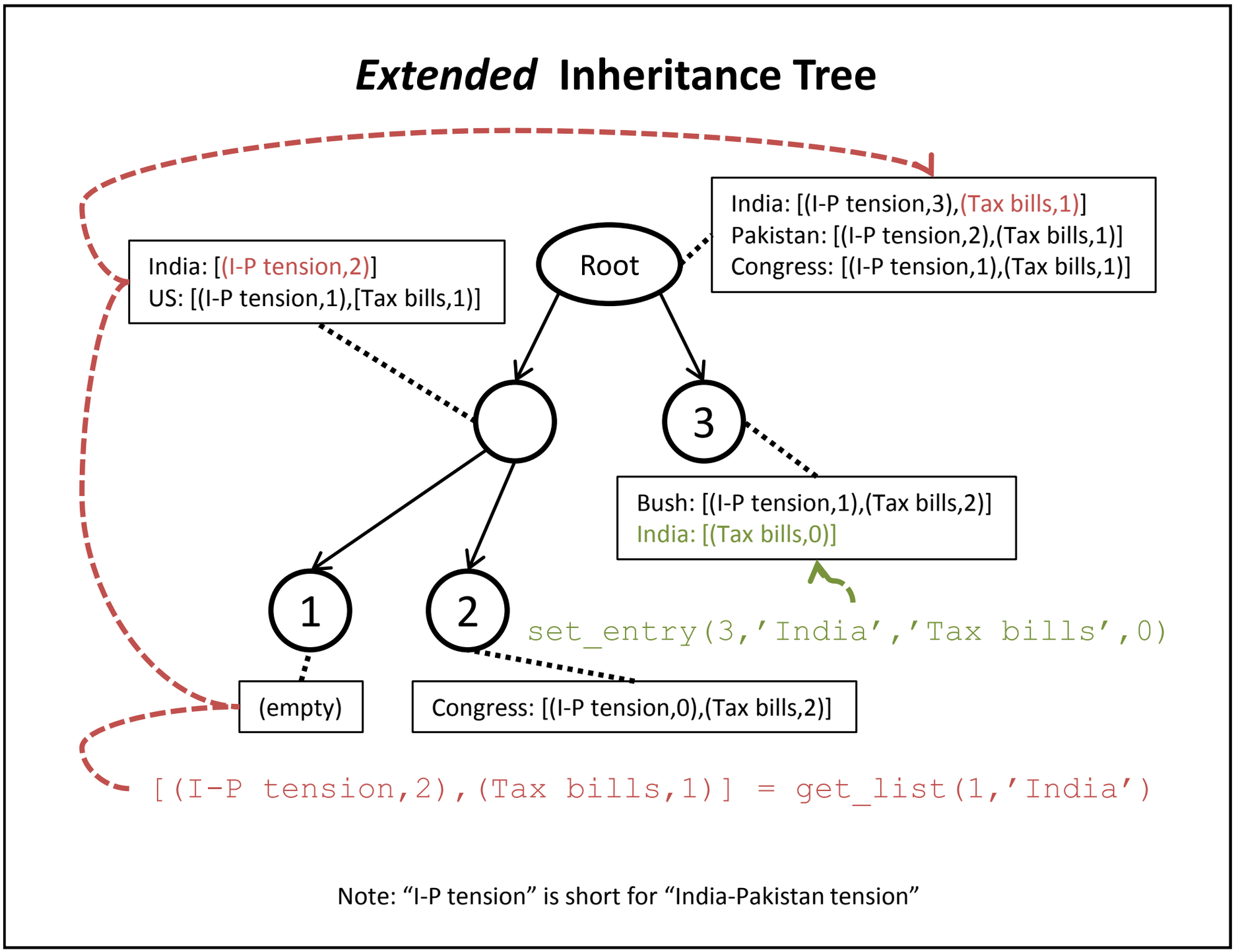
past state

- Resample particles if l_2 norm too large
  (resample some assignments for diversity, too)
- Compare to multiplicative updates algorithm
  In our case predictive likelihood yields weights

Inheritance Tree



Extended Inheritance Tree



write only in the leaves (per thread)

Results

Numbers …

- TDT5 (Topic Detection and Tracking)
  macro-averaged minimum detection cost: 0.714

time	entities	topics	story words
0.84	0.90	0.86	0.75

  This is the best performance on TDT5!

- Yahoo News data
  … beats all other clustering algorithms

Stories

TOPICS

Sports

games
won
team
final
season
league
held

Politics

government
minister
authorities
opposition
officials
leaders
group

Unrest

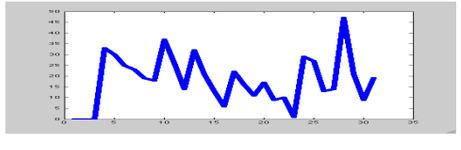
police
attack
run
man
group
arrested
move

STORYLINES

UEFA-soccer

champions
goal
leg
coach
striker
midfield
penalty

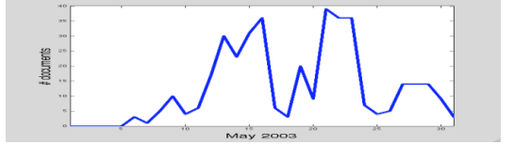
Juventus
AC Milan
Real Madrid
Milan
Lazio
Ronaldo
Lyon



Tax bills

tax
billion
cut
plan
budget
economy
lawmakers

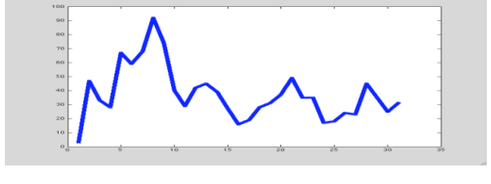
Bush
Senate
US
Congress
Fleischer
White House
Republican



India-Pakistan tension

nuclear
border
dialogue
diplomatic
militant
insurgency
missile

Pakistan
India
Kashmir
New Delhi
Islamabad
Musharraf
Vajpayee



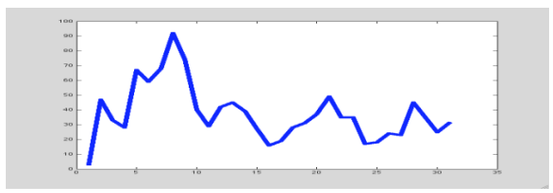
Related Stories

"Show similar stories by topic"

"Show similar stories, require the word nuclear*"*

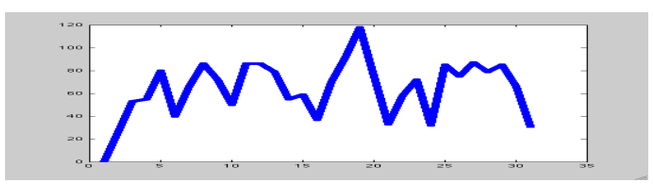
India-Pakistan tension

nuclear	*Pakistan*
border	*India*
dialogue	*Kashmir*
diplomatic	*New Delhi*
militant	*Islamabad*
insurgency	*Musharraf*
missile	*Vajpayee*



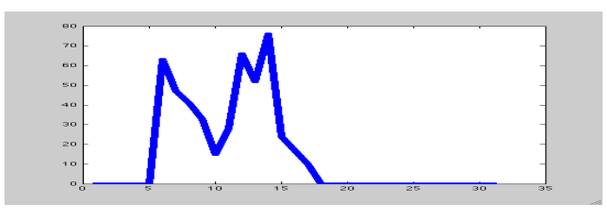
Middle-east conflict

Peace	*Israel*
Roadmap	*Palestinian*
Suicide	*West bank*
Violence	*Sharon*
Settlements	*Hamas*
bombing	*Arafat*



North Korea nuclear

nuclear	*North Korea*
summit	*South Korea*
warning	*U.S*
policy	*Bush*
missile	*Pyongyang*
program	



Issues

To Do