Bursty and Hierarchical Structure in Streams

Jon Kleinberg

Cornell University

Topics and Time

Documents can be organized by topic,

but we also experience their arrival over time.

- **E-mail, news articles.**
- **Research papers, on a slower time scale.**

(1) Temporal sub-structure within a single topic.

(Nested) bursts of activity surrounding events.

(2) Time-line construction: enumeration of topics over time.

[Allen 1995, Kumar et al. 1997, Swan-Allan 2000, Swan-Jensen 2000]

[Topic Detection and Tracking: Allan et al. 1998, Yang et al. 1998]

Develop techniques based on Markov source models for

temporal text mining.

Mining E-mail

E-mail archives as a domain for data mining.

- **Raw material for historical research and legal proceedings.**
  (Natl. Archives: >10 million e-mail msgs from Clinton White House)
- **Personal archives can reach 10-100's MB of pure text.**

Topic-based organization (automated folder management):

[Helfman-Isbell 95, Cohen 96, Lewis-Knowles 97, Sahami et al. 98, Segal-Kephart 99, Horvitz 99, Rennie 00]

Flow of time exposes sub-structure in a coherent folder

For example, folder on "grant proposals" contains multiple bursty periods corresponding to localized episodes.

E.g. "the process of gathering people for our large NSF ITR proposal."

The role of time in narratives

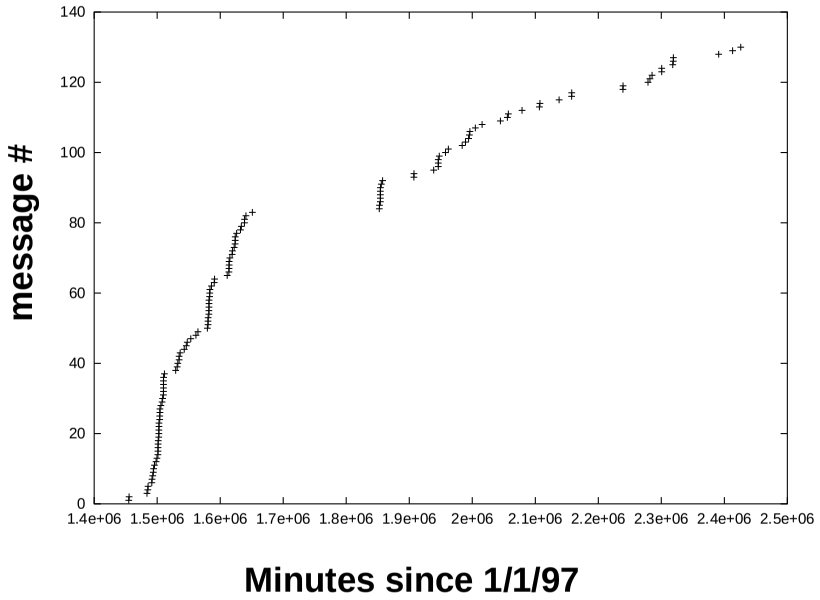
. . . there seems something else in life besides time, something which may conveniently be called "value," something which is measured not by minutes or hours but by intensity, so that when we look at our past it does not stretch back evenly but piles up into a few notable pinnacles, and when we look at the future it seems sometimes a wall, sometimes a cloud, sometimes a sun, but never a chronological chart.

- E.M. Forster, Aspects of the Novel (1928)

- **Anisochronies in narratives [Genette 1980, Chatman 1978]:**
  non-uniform relation between time span of a story's events and the time it takes to relate them.

Intensity? Notable Pinnacles?

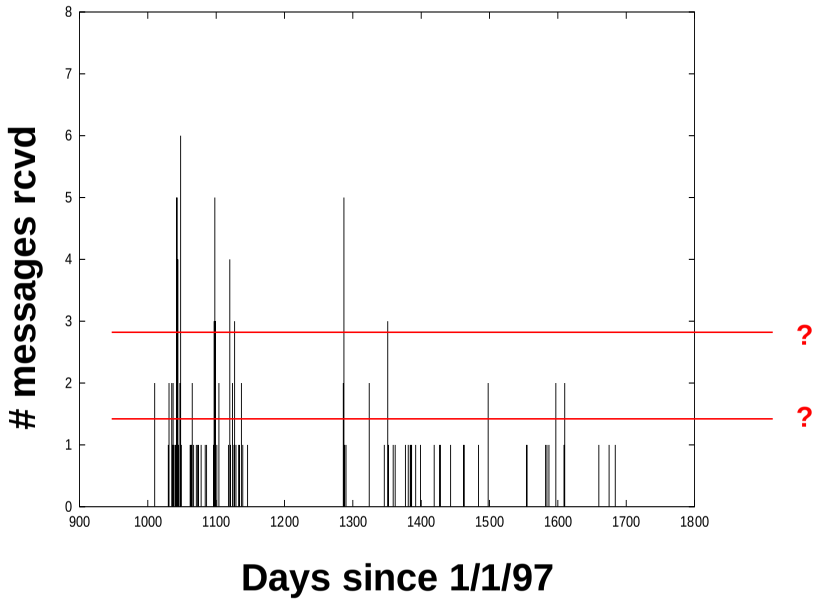
"I know a burst when I see one." $\longrightarrow$??



Need a precise model:

- **Inspection not likely to give the full structure in the sequence.**
- **Eventually want to perform burst detection for all terms in corpus.**

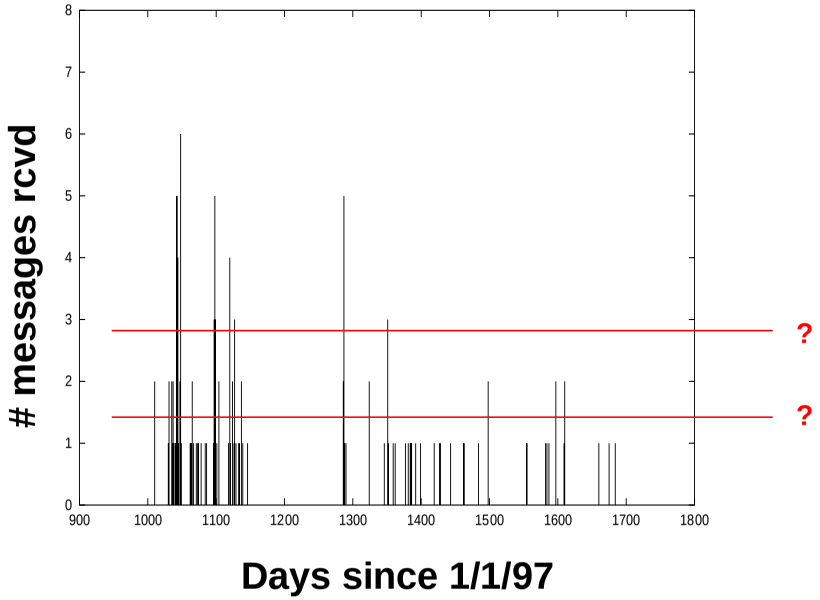
Threshold-Based Methods



Swan-Allan [1999, 2000], Swan-Jensen [2000] introduced threshold-based methods.

- **Bin relevant messages by day.**
- **Identify days in which number of relevant messages is above a computed threshold (χ^2 or similar test).**
- **Contiguous set of days above threshold constitutes an episode.**

Threshold-Based Methods



Issues for threshold-based methods as a baseline:

- **E-mail folders quite sparse/noisy.**

  E.g. in figure, no 7 consecutive days with non-zero # of messages.

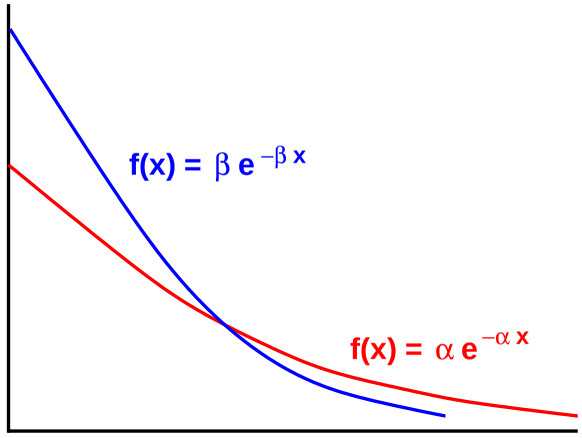
- **We want to find episodes lasting several months**

  (e.g. writing a proposal) as well as several days.

  Multiple time scales? Bursts within bursts?

A Model for Bursty Streams

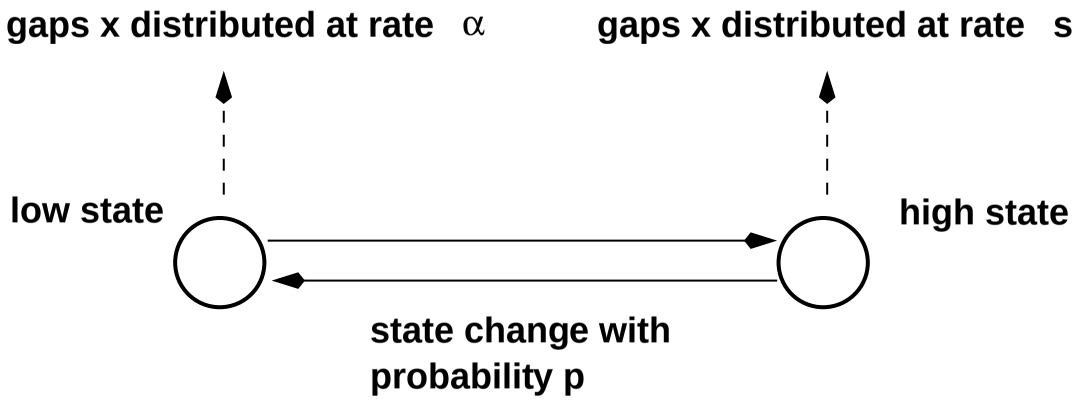
Want a source model for messages, determining arrival times.



Simplest: exponential distribution.

- **Gap in time x until next message is distributed according to $f(x) = \alpha e^{-\alpha x}$. ("Memoryless" distribution.)**
- **Expected gap value is α^{-1}. Thus α is called the "rate" of message arrivals.**

A Model for Bursty Streams



A model for message generation with persistent bursts:

Markov source model [e.g. Anick-Mitra-Sondhi 1982, Scott 1998]

- **Low state q_0: gaps in time between message arrivals distributed according to exponential distribution with rate α.**
- **High state q_1: gaps distributed at rate $s\alpha$, where $s > 1$.**
- **Before each message emission, state changes with probability p.**

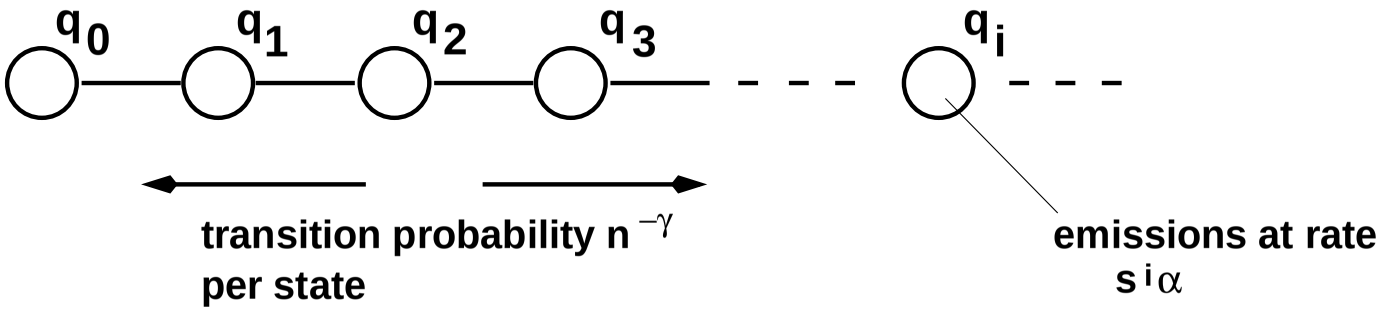
Consider $n + 1$ messages, with positive gaps between arrival times.

Most likely state sequence via Bayes' Thm and dynamic programming.

A Richer Model

Want to model bursts of greater and greater intensity

$\longrightarrow$ set of states representing arbitrarily small gap sizes.



Infinite state set $q_0, q_1, q_2, \ldots$

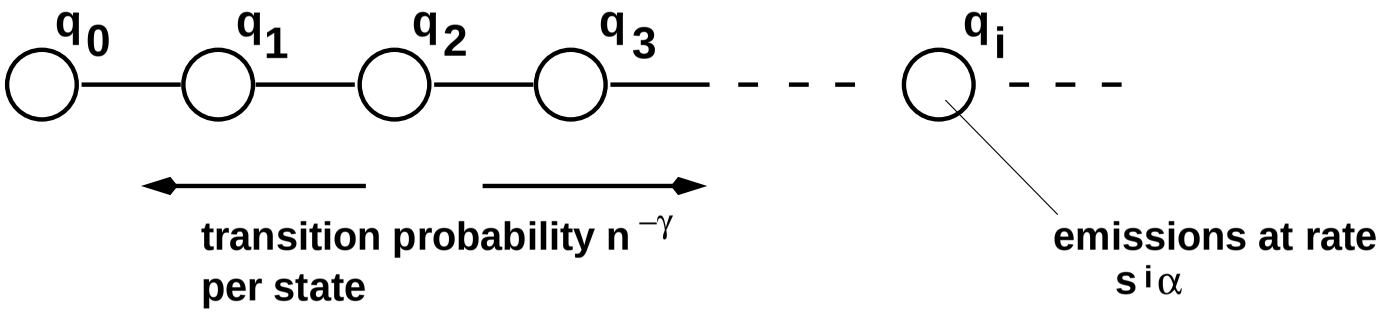
- **If** n **gaps over time** T**, then average rate** $\alpha = n/T$**.**

  $\longrightarrow$ **"base rate" at** q_0 **is** α**.**

- **Rates increase by factor of** s**: rate for** q_i **is** $s^i\alpha$**.**

- **Jumping from** q_i **to** q_j **in one step has prob.** $n^{-|j-i|\gamma}$**.**

A Richer Model



Theorem: Let $\delta(\mathbf{x}) = \min_{i=1}^{n} x_i$**.**

The maximum likelihood state sequence involves only states $q_0, q_1, \ldots, q_k$**, where** $k \leq \lceil 1 + \log_s T + \log_s \delta(\mathbf{x}) \rceil$**.**

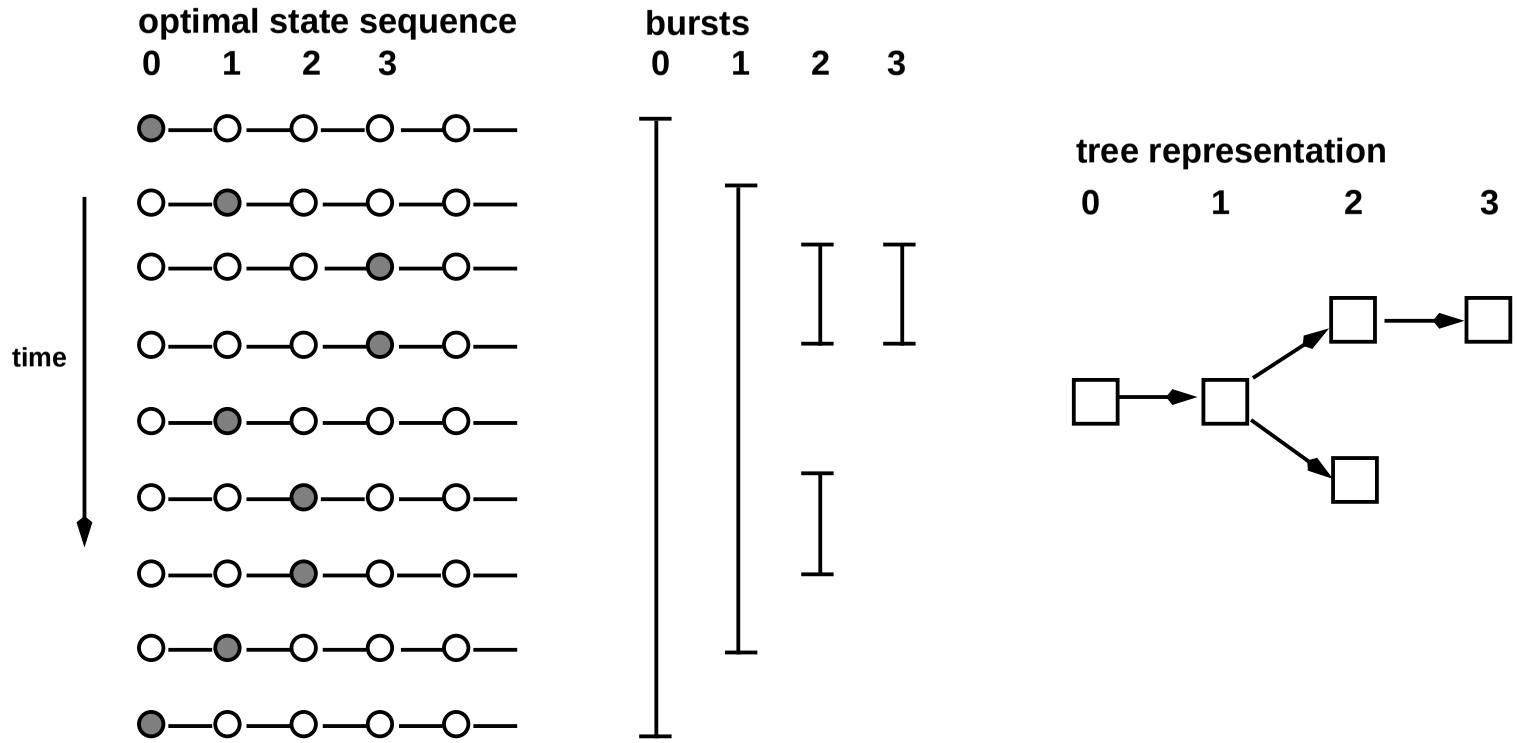
Using Theorem, can reduce to the finite-state case and apply dynamic programming.

(Cf. Viterbi algorithm for Hidden Markov models.)

Hierarchical Structure

Define a burst of intensity j to be a maximal interval in which optimal state sequence is in state q_j or higher.

Bursts are naturally nested: each burst of intensity j is contained in a unique burst of intensity $j-1$ $\longrightarrow$ hierarchical tree structure.



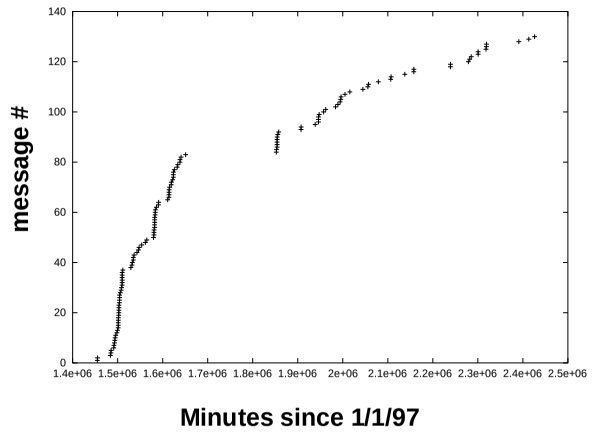
Experiments with an E-Mail Stream

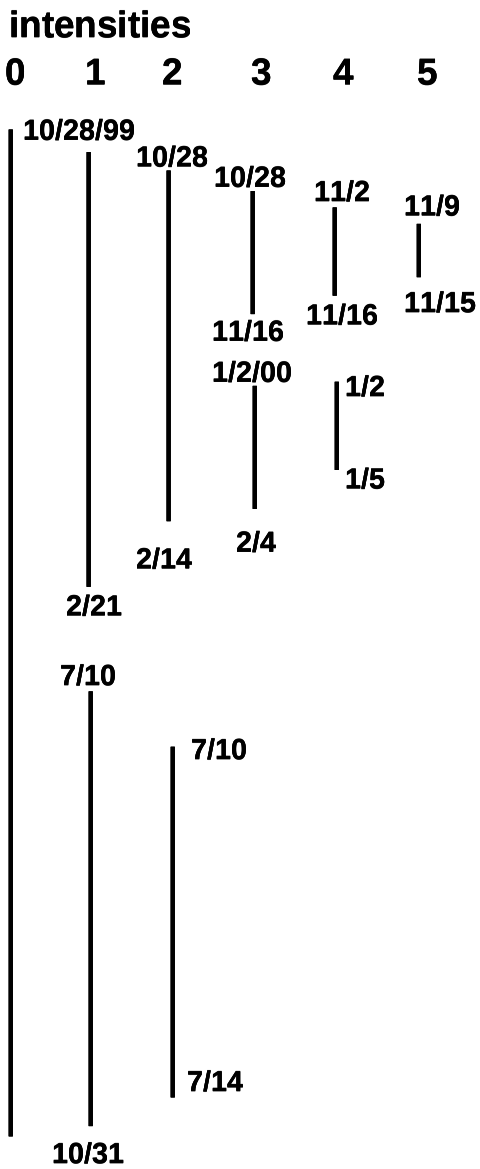
As a proxy for folders, look at queries to e-mail archive.

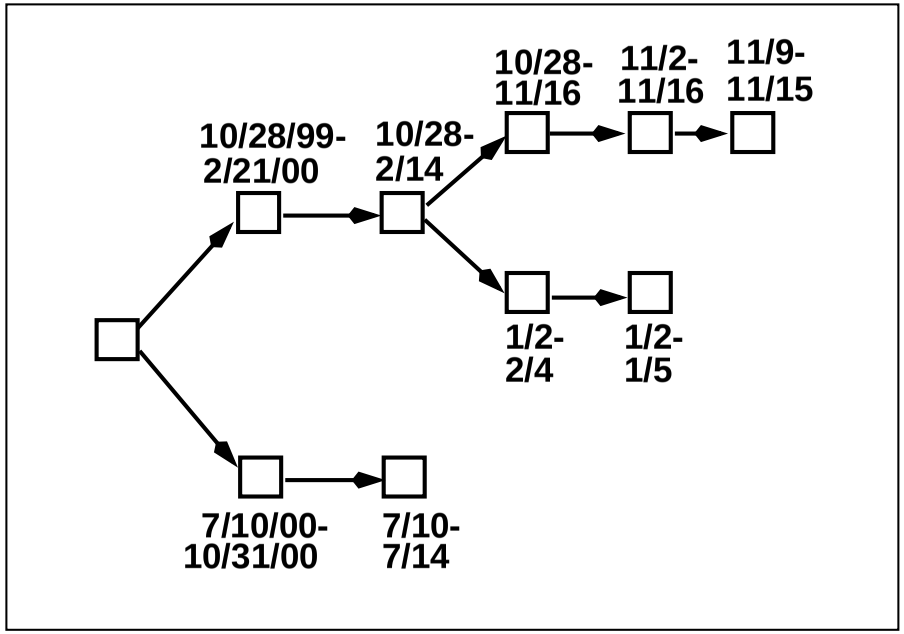
- **Simple implementation of algorithm can build burst representation for a query in real-time.**
- **Do spikes emerge in vicinity of recognizable events?**

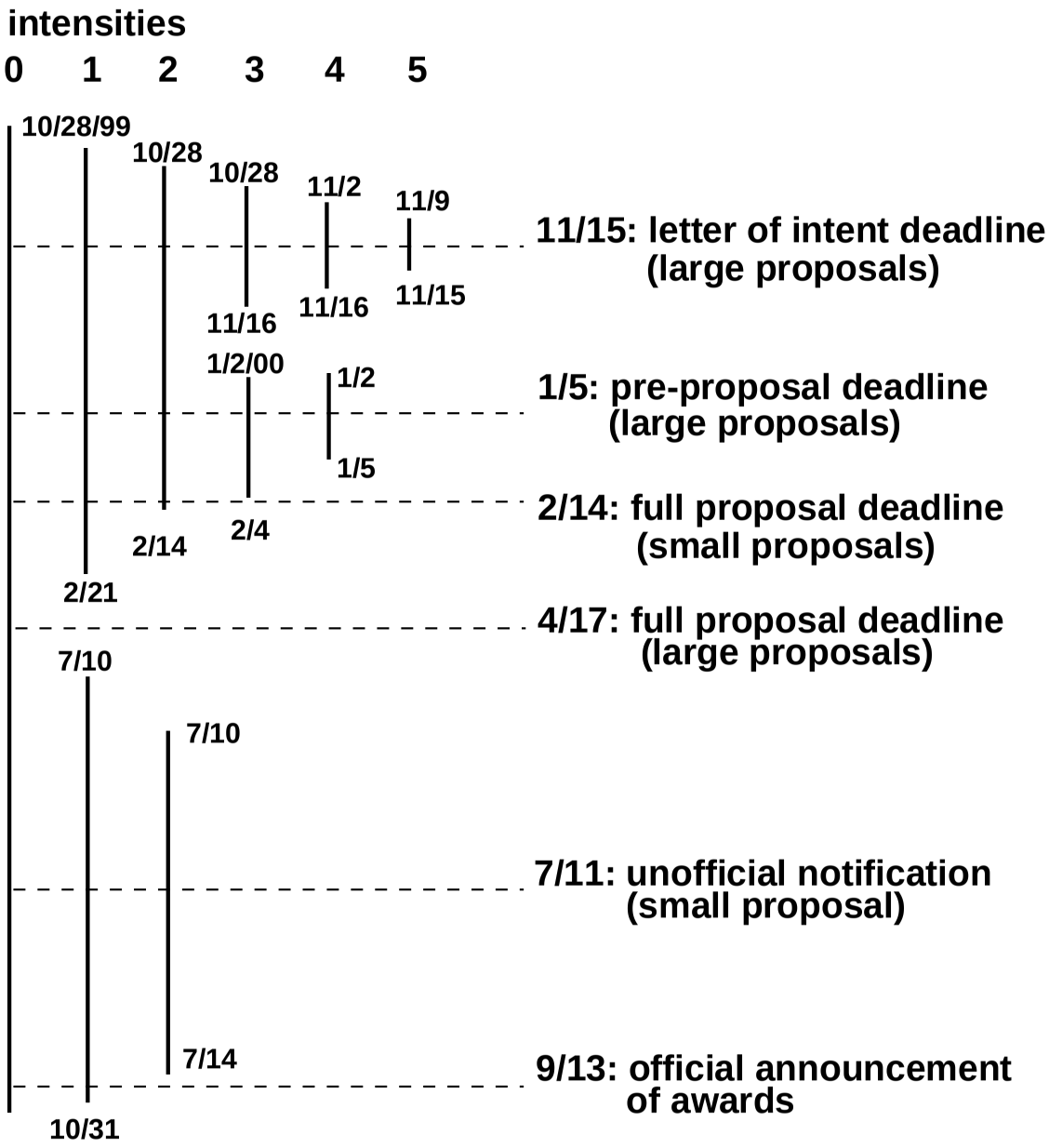
Example: stream of all messages containing the word "ITR."

(Large NSF program; applied for two proposals (large and small) with colleagues in academic year 1999-2000.)




intensities
0
1
2
3
4
5
10/28/99
10/28
10/28
11/2
11/9
11/16
11/16
11/15
1/2/00
1/2
1/5
2/4
2/14
2/21
7/10
7/10
7/14
10/31


10/28-
11/16
11/2-
11/16
11/9-
11/15
10/28/99-
2/21/00
10/28-
2/14
1/2-
2/4
1/2-
1/5
7/10/00-
10/31/00
7/10-
7/14


intensities
0
1
2
3
4
5
10/28/99
2/21
10/28
2/14
10/28
11/16
11/2
11/16
11/9
11/15
11/15: letter of intent deadline
(large proposals)
1/2/00
2/4
1/2
1/5
1/5: pre-proposal deadline
(large proposals)
2/14: full proposal deadline
(small proposals)
4/17: full proposal deadline
(large proposals)
7/10
10/31
7/10
7/14
7/11: unofficial notification
(small proposal)
9/13: official announcement
of awards

Query: “Prelim”

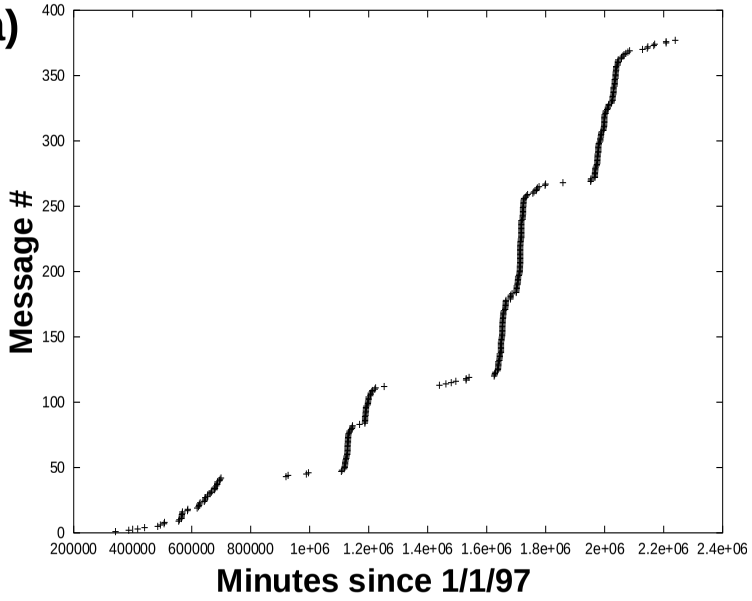
Example: stream of all messages containing the word “prelim.”

(Cornell terminology for a non-final exam in an undergraduate course.)

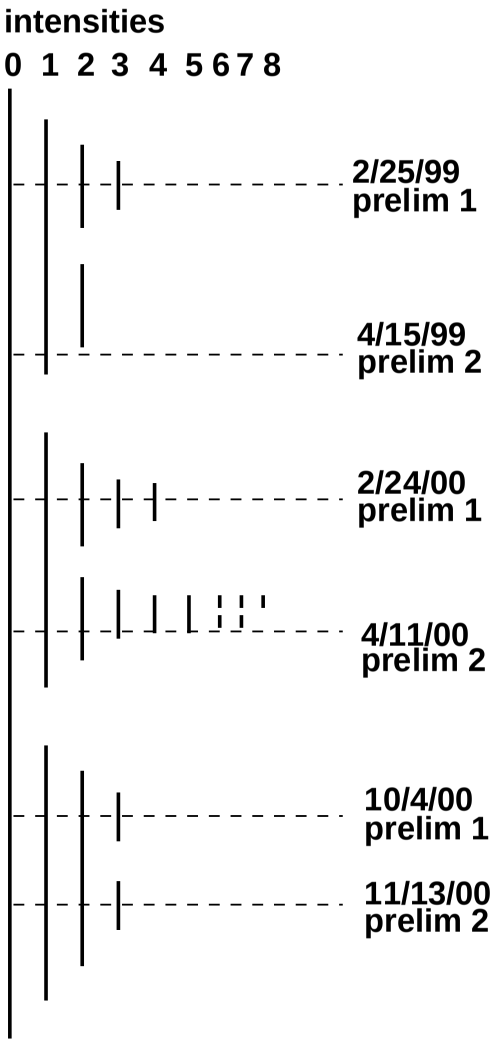
- **E-mail archive spans four large courses, each with two prelims.**
- **But in first course, almost all correspondence restricted to course e-mail account.**

Three large courses, two prelims in each.

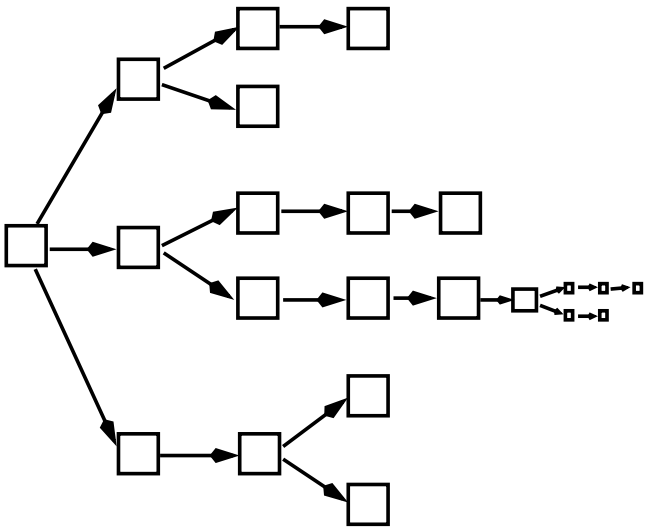
a)


Message #
0
50
100
150
200
250
300
350
400
200000
400000
600000
800000
1e+06
1.2e+06
1.4e+06
1.6e+06
1.8e+06
2e+06
2.2e+06
2.4e+06
Minutes since 1/1/97

b)


intensities
0 1 2 3 4 5 6 7 8
2/25/99
prelim 1
4/15/99
prelim 2
2/24/00
prelim 1
4/11/00
prelim 2
10/4/00
prelim 1
11/13/00
prelim 2

c)



Enumerating Bursts for Time-Line Construction

Can enumerate bursts for every word in the corpus.

- **Essentially one pass over an inverted index.**
- **Weight of burst B of intensity $j = \log(\Pr[q_j \mid B] / \Pr[q_{j-1} \mid B])$.**

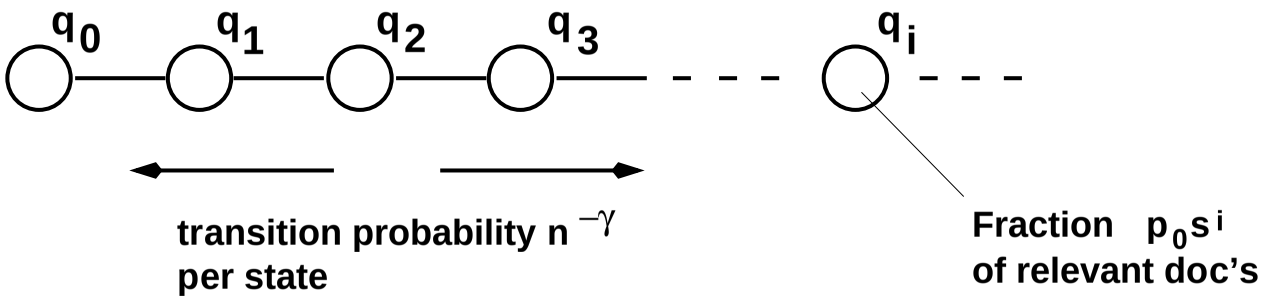
Over history of a conference or journal, topics rise/fall in significance.

Using words as stand-ins for topic labels:

What are the most prominent topics at different points in time?

- **Take words in paper titles over history of conference.**
- **Compute bursts for each word; find those of greatest weight.**
- **All words are considered. (Even stop-words.)**

A Source Model for Batched Arrivals

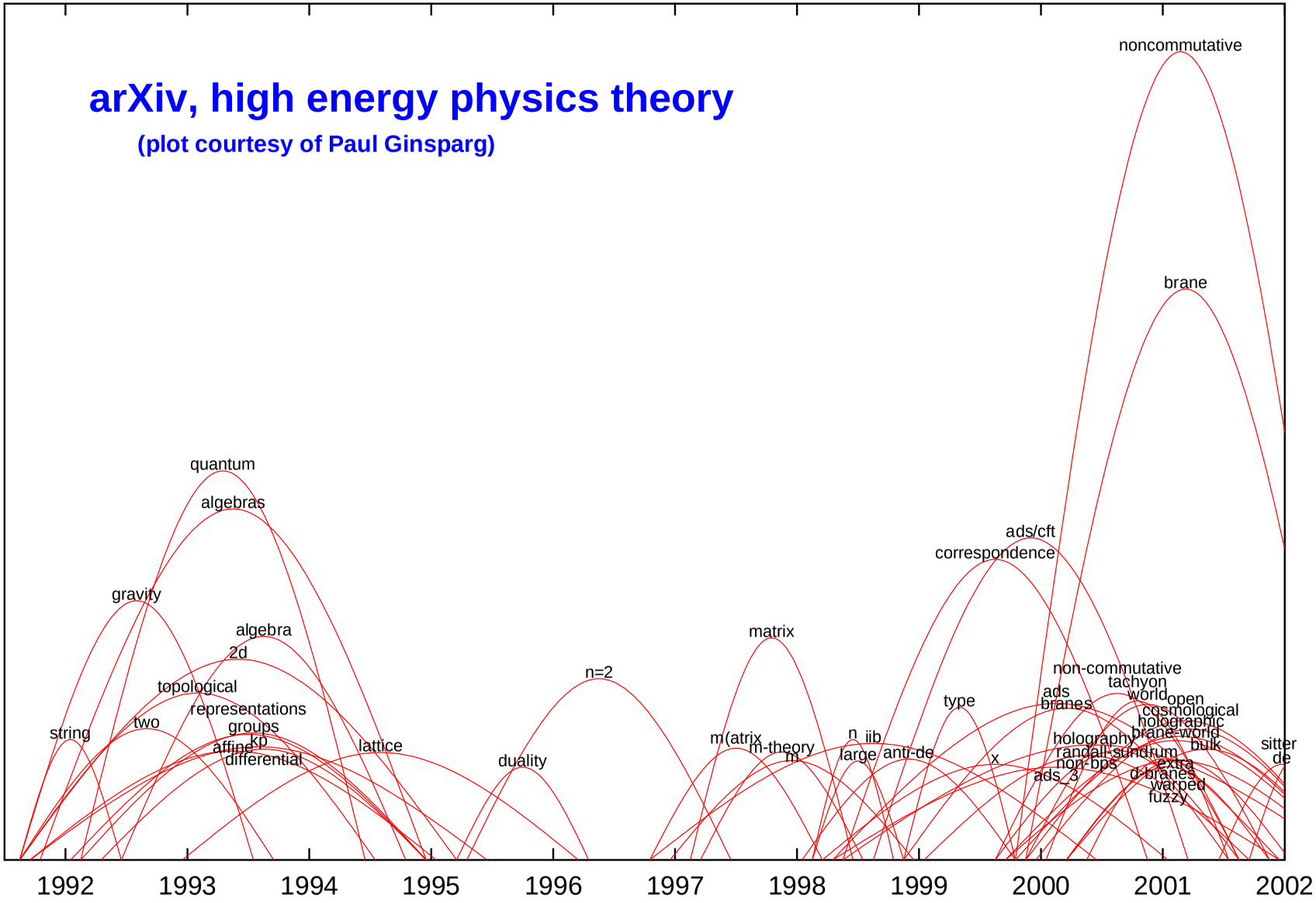


- n **batches of documents. Batch** t **contains** d_t **total, of which** r_t **are relevant (e.g. contain fixed word).**
- **Overall relevant fraction** $p_0 = (\sum r_t)/(\sum d_t)$**.**
- **State** q_i**: expected fraction of relevant documents** $p_i = p_0 s^i$**.**

Word	Interval of burst		
grammars	1969 STOC — 1973 FOCS	logic	1976 FOCS — 1984 STOC
automata	1969 STOC — 1974 STOC	vlsi	1980 FOCS — 1986 STOC
languages	1969 STOC — 1977 STOC	probabilistic	1981 FOCS — 1986 FOCS
machines	1969 STOC — 1978 STOC	how	1982 STOC — 1988 STOC
recursive	1969 STOC — 1979 FOCS	parallel	1984 STOC — 1987 FOCS
classes	1969 STOC — 1981 FOCS	algorithm	1984 FOCS — 1987 FOCS
some	1969 STOC — 1980 FOCS	graphs	1987 STOC — 1989 STOC
sequential	1969 FOCS — 1972 FOCS	learning	1987 FOCS — 1997 FOCS
equivalence	1969 FOCS — 1981 FOCS	competitive	1990 FOCS — 1994 FOCS
programs	1969 FOCS — 1986 FOCS	randomized	1992 STOC — 1995 STOC
program	1970 FOCS — 1978 STOC	approximation	1993 STOC —
on	1973 FOCS — 1976 STOC	improved	1994 STOC — 2000 STOC
complexity	1974 STOC — 1975 FOCS	codes	1994 FOCS —
problems	1975 FOCS — 1976 FOCS	approximating	1995 FOCS —
relational	1975 FOCS — 1982 FOCS	quantum	1996 FOCS —

Word	Interval of burst
data	1975 SIGMD — 1979 SIGMD
base	1975 SIGMD — 1981 VLDB
application	1975 SIGMD — 1982 SIGMD
bases	1975 SIGMD — 1982 VLDB
design	1975 SIGMD — 1985 VLDB
relational	1975 SIGMD — 1989 VLDB
model	1975 SIGMD — 1992 VLDB
large	1975 VLDB — 1977 VLDB
schema	1975 VLDB — 1980 VLDB
theory	1977 VLDB — 1984 SIGMD
distributed	1977 VLDB — 1985 SIGMD
data	1980 VLDB — 1981 VLDB
statistical	1981 VLDB — 1984 VLDB
database	1982 SIGMD — 1987 VLDB
nested	1984 VLDB — 1991 VLDB

deductive	1985 VLDB — 1994 VLDB
transaction	1987 SIGMD — 1992 SIGMD
objects	1987 VLDB — 1992 SIGMD
object-oriented	1987 SIGMD — 1994 VLDB
parallel	1989 VLDB — 1996 VLDB
object	1990 SIGMD — 1996 VLDB
mining	1995 VLDB —
server	1996 SIGMD — 2000 VLDB
sql	1996 VLDB — 2000 VLDB
warehouse	1996 VLDB —
similarity	1997 SIGMD —
approximate	1997 VLDB —
web	1998 SIGMD —
indexing	1999 SIGMD —
xml	1999 VLDB —


arXiv, high energy physics theory
(plot courtesy of Paul Ginsparg)
noncommutative
brane
quantum
algebras
gravity
algebra
2d
topological
representations
groups
kp
affine
differential
two
string
lattice
n=2
duality
matrix
m(atrix
m-theory
m
ads/cft
correspondence
type
x
n
iib
large
anti-de
non-commutative
tachyon
ads
world
branes
open
cosmological
holographic
brane-world
holography
bulk
randall-sundrum
non-bps
extra
ads_3
d-branes
warped
fuzzy
sitter
de
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002

Some Observations

Many of the bursts contain significant number of batches with few/no relevant documents. (cf. threshold-based methods.)

Words with highest-weight bursts different from most frequent words.

- **Most frequent words in STOC/FOCS titles:**

  of, for, the, and, a, on, in, complexity, algorithms, with, to, problems, time, parallel, algorithm, bounds, problem, graphs, an, lower

- **Bursty words almost always content-bearing.**
  But content-bearing words not always bursty.
  E.g. "time" and "bounds" common throughout all years.

- **Burst weight represents balance between ubiquity and abruptness.**

- **Relative rates of high and low states (parameter s) determines whether we find brief, intense bursts or longer, milder bursts.**

Word	Interval of burst		
depression	1930 – 1937	collective	1947 – 1961
recovery	1930 – 1937	aggression	1949 – 1955
banks	1931 – 1934	defense	1951 – 1952
democracy	1937 – 1941	free	1951 – 1953
wartime	1941 – 1947	soviet	1951 – 1953
production	1942 – 1943	korea	1951 – 1954
fighting	1942 – 1945	communist	1951 – 1958
japanese	1942 – 1945	program	1954 – 1956
war	1942 – 1945	alliance	1961 – 1966
peacetime	1945 – 1947	communist	1961 – 1967
program	1946 – 1948	poverty	1963 – 1969
veterans	1946 – 1948	propose	1965 – 1968
wage	1946 – 1949	tonight	1965 – 1969
housing	1946 – 1950	billion	1966 – 1969
atomic	1947 – 1959	vietnam	1966 – 1973

Some Observations

Is it the content that's bursty, or just the time series?

Permutation test (see [Swan-Jensen 2000])

- **Start with full e-mail corpus, arrival times $t_1, \ldots, t_N$.**
- **Shuffle messages via random permutation π:**
  message $\pi(i)$ arrives at time t_i (instead of message i).
- **Total weight of all bursts in shuffled corpus more than order of magnitude smaller than in true corpus (25K vs. 370K)**
- **Almost no hierarchy in shuffled version: average of 16 words with depth ≥ 2, versus 3865 in true corpus.**

Further Related Work

Markov source models for time-series analysis

- **Fraud detection, Web page requests [Scott 98, Scott-Smyth 02].**

Piece-wise function approximation

- **Long history in statistics [Hudson 1966, Hawkins 1976].**
- **Recent applications in data mining for trend and event detection [Keogh-Smyth 1997, Han et al. 1998, Mannila-Salmenkivi 2001]**

Constructing trees from time series

- **Waveform $\longrightarrow$ branches at local minima, leaves at local maxima. [Ehrich-Foith 1976, Shaw-DeFigueiredo 1990]**
- **Hierarchical HMMs [Fine-Singer-Tishby 1998, Murphy-Paskin 2001]**

Visualization of news streams

- **Wavelet Analysis [Miller et al. 98], ThemeRiver [Havre et al. 2000].**

Further Directions

Web clickstream data

- **Logs collected by Gay, Stefanone, Grace-Martin, Hembrooke 2000.**
- **80 undergraduates in two classes, early March to mid-May 2000, with consent.**
- **Bursts correspond to sudden rise in site traffic.**
  Great difference between single-user bursts and bursts involving more than e.g. 10 distinct users.
- **Many of the heaviest multi-user bursts involve URLs of on-line class reading assignments, just before and during discussion section.**

Similar domains:

- **Search engine query logs. (cf. Google Zeitgeist)**
- **Superposition of downloading and paper submission in the arXiv.**

Open Questions

Data stream computation

- **In a data stream model, find bursts of large weight for all items (e.g. all possible words) simultaneously.**
- **One pass, limited storage.**

On-line algorithms

- **Given a stream of e-mail messages / paper titles / paper downloads, how early, in an on-line setting, can a large-weight burst be identified?**
- **Detecting the emergence of significant new topics as they happen.**
  (cf. first-story detection problem in TDT).

Reflections

The fact that we need tools to pre-screen our email for us just shows how information-overloaded our society has become.

– Slashdot posting

24 April 2002, 2:10 PM

Who the @#$! gets so much email they need to mine for text ??!! dont change your email filtering, change your pathetic life !!

– Slashdot posting

24 April 2002, 6:02 PM

If only it were so simple ...

- **Increasingly able to measure personal activity at unprecedented levels of detail.**
- **Coping with a world in which your on-line tools know more about you than you realize.**