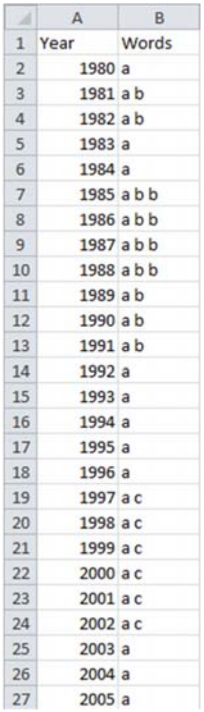
Burst Detection

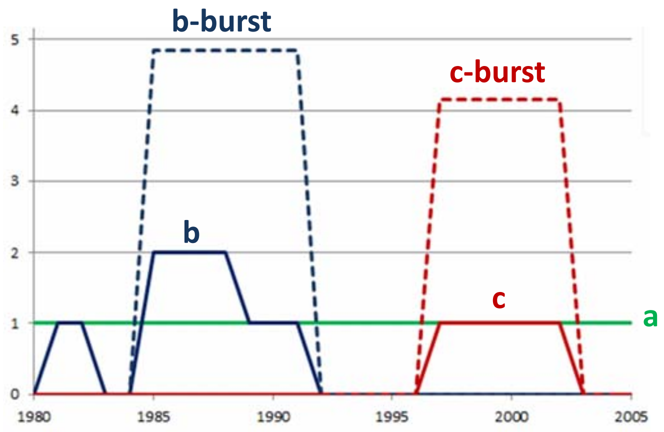


	A	B
1	Year	Words
2	1980	a
3	1981	a b
4	1982	a b
5	1983	a
6	1984	a
7	1985	a b b
8	1986	a b b
9	1987	a b b
10	1988	a b b
11	1989	a b
12	1990	a b
13	1991	a b
14	1992	a
15	1993	a
16	1994	a
17	1995	a
18	1996	a
19	1997	a c
20	1998	a c
21	1999	a c
22	2000	a c
23	2001	a c
24	2002	a c
25	2003	a
26	2004	a
27	2005	a

Kleinberg's burst-detection algorithm identifies sudden increases in the frequency of words.

Given time-stamped text, it identifies words that burst.

“a” does not burst. “b” bursts more than “c.”



Information Visualization MOOC

Unit 2 – “When”: Temporal Data

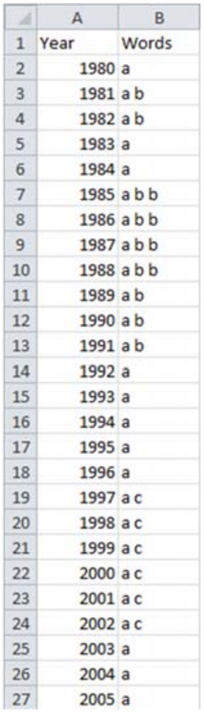
Burst Detection

Kleinberg, Jon M. 1998. “Authoritative Sources in a Hyperlinked Environment.” *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms,* 668–677.

Relevant Research Disciplines:
Mathematics, Statistics, Information Visualization

CNS Cyberinfrastructure for Network Science Center http://ivmooc.cns.iu.edu INDIANA UNIVERSITY

Burst Detection

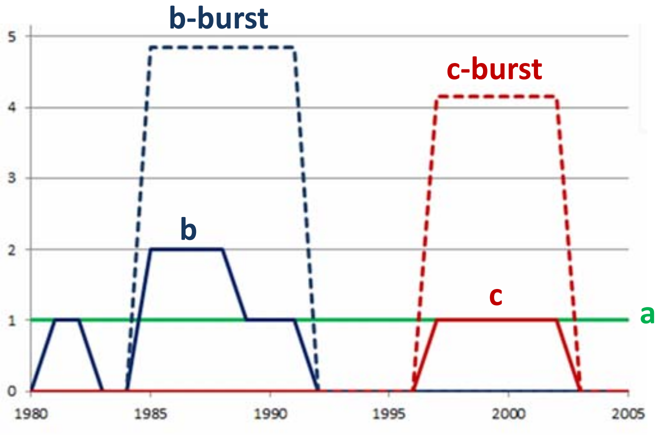


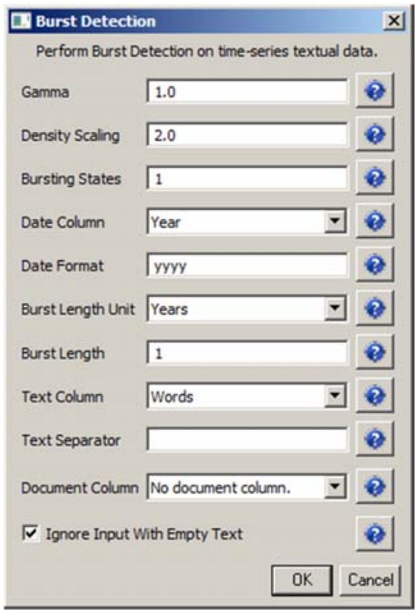
	A	B
1	Year	Words
2	1980	a
3	1981	a b
4	1982	a b
5	1983	a
6	1984	a
7	1985	a b b
8	1986	a b b
9	1987	a b b
10	1988	a b b
11	1989	a b
12	1990	a b
13	1991	a b
14	1992	a
15	1993	a
16	1994	a
17	1995	a
18	1996	a
19	1997	a c
20	1998	a c
21	1999	a c
22	2000	a c
23	2001	a c
24	2002	a c
25	2003	a
26	2004	a
27	2005	a

Kleinberg's burst-detection algorithm identifies sudden increases in the frequency of words.

Given time-stamped text, it identifies words that burst.

“a” does not burst. “b” bursts more than “c.”





Perform Burst Detection on time-series textual data.

Gamma: 1.0

Density Scaling: 2.0

Bursting States: 1

Date Column: Year

Date Format: yyyy

Burst Length Unit: Years

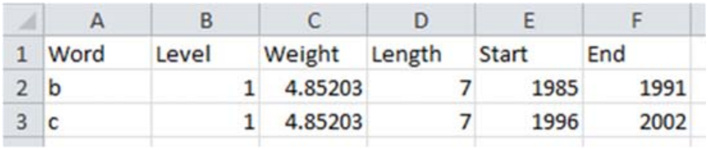
Burst Length: 1

Text Column: Words

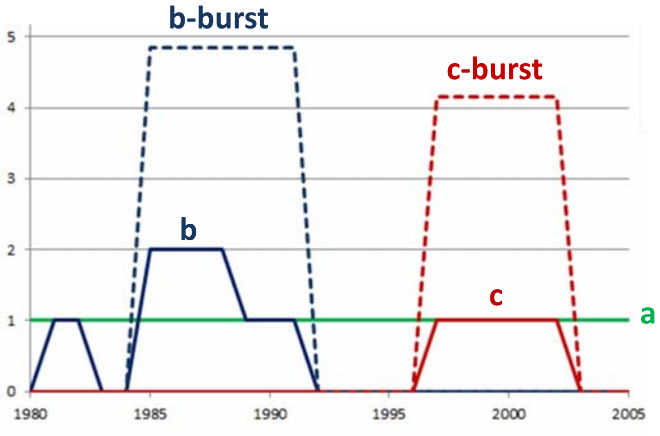
Text Separator:

Document Column: No document column.

☑ Ignore Input With Empty Text



	A	B	C	D	E	F
1	Word	Level	Weight	Length	Start	End
2	b	1	4.85203	7	1985	1991
3	c	1	4.85203	7	1996	2002



‘Text Separator’ is a space.

Text Normalization (see Topical Analysis)

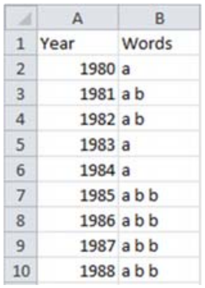
Sample text: “Emergence of Scaling in Random Networks”

- Lowercase: The example text becomes "emergence of scaling in random networks.”
- Tokenize: The text blob is split into a list of individual words. The example text becomes "emergence|of|scaling|in|random|networks.”
- Stem: Common or low-content prefixes and suffixes are removed to identify the core concept. The example text becomes "emerg|of|scale|in|random|network.”
- Stopword: Low-content tokens like "of" and "in" are removed (see the complete stopword list). The example text becomes "emerg|scale|random|network.”

Burst-Detection Algorithm

Given:

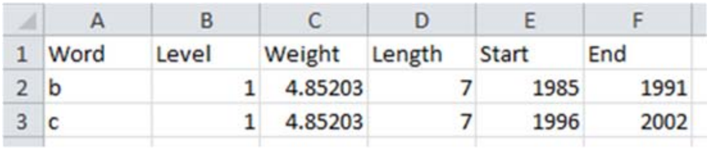
- A stream of events
- Every event = set of keywords + time stamp



	A	B
1	Year	Words
2	1980	a
3	1981	a b
4	1982	a b
5	1983	a
6	1984	a
7	1985	a b b
8	1986	a b b
9	1987	a b b
10	1988	a b b

Identify:

- Time intervals with unusually high frequency of a specific keyword.



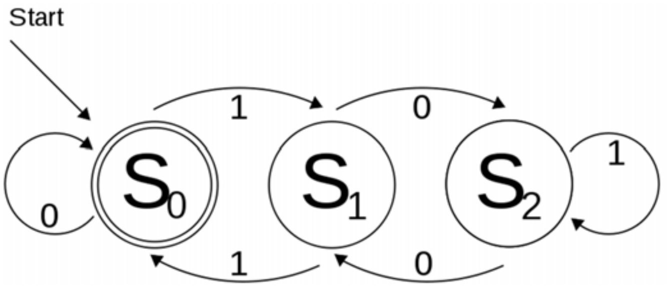
	A	B	C	D	E	F
1	Word	Level	Weight	Length	Start	End
2	b	1	4.85203	7	1985	1991
3	c	1	4.85203	7	1996	2002

Burst-Detection Algorithm

Applies Hidden Markov Model—it is assumed that:

- Imaginary finite automaton generates event stream.
- Finite automaton has known structure but unknown state sequence.
- Find optimal state sequence of states that best fits data.
- Bursts are states of the imaginary automaton—if sequence is known, all bursts are known.

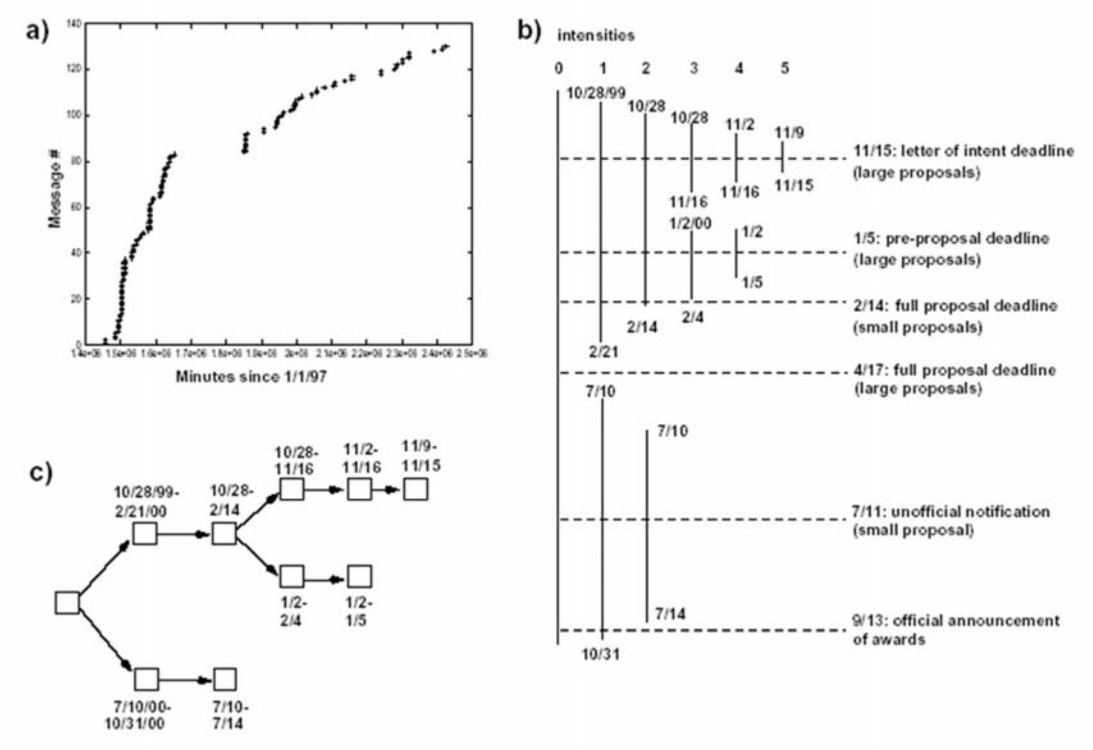
sample
state diagram



Burst-Detection Algorithm

Compute state sequence using dynamic programming:

- For every day *d* and every state *s,* compute optimal state sequence for period [1..*d*] ending with state *s*.
- Given data for the next day, try all values for yesterday and choose the best one.
- For optimal sequence for the whole interval [1..D], take maximum over all states.



Applications

Are there spikes of activity in email, Twitter, Flickr, or news data streams—e.g., due to external events, deadlines?

Are there surges of interest—e.g.,

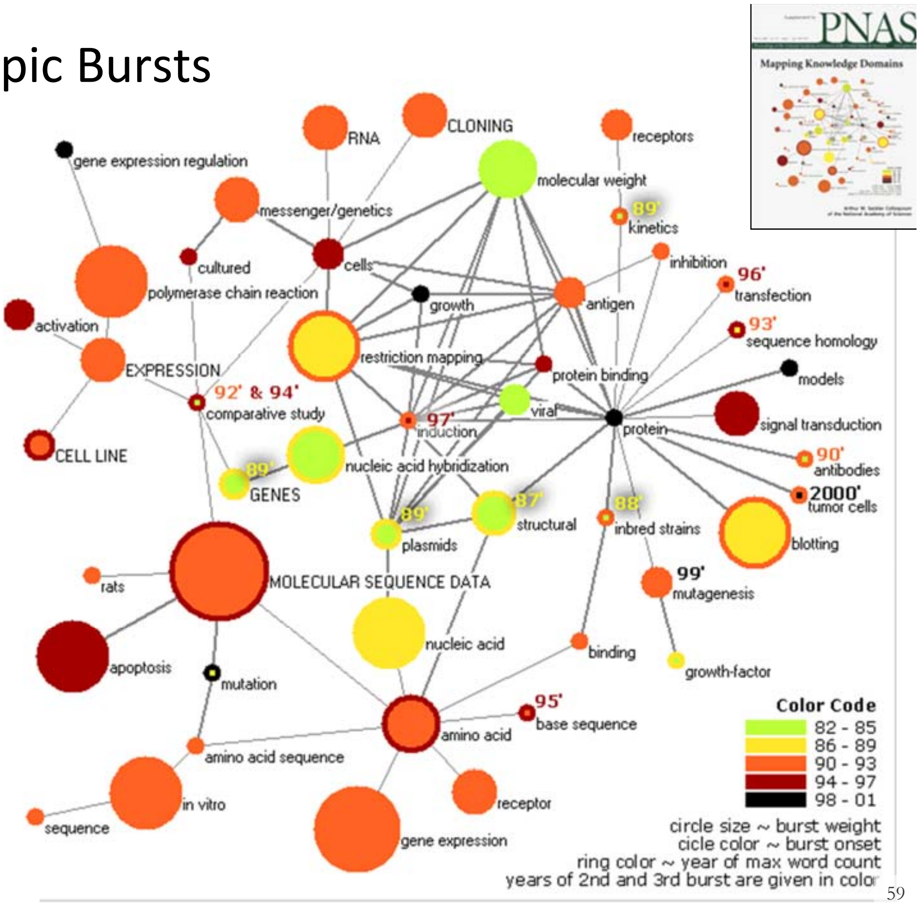
- Bursts in the download activity of certain files,
- Number of citations to specific papers,
- Funding awarded to an institution, or
- The number of new friends/collaborators a person/company has over time?

Are today’s bursts of activity an indicator for tomorrow’s adoption or high usage of a word or idea?

Mapping Topic Bursts

Co-word space of the top-50 most frequent and bursty words used in the top-10% most highly cited *PNAS* publications in 1982-2001.

Mane & Börner. 2004. PNAS 101(Suppl. 1): 5287-5290.



Scalability & Known Issues

The algorithms are efficient enough that computing a representation for the bursts on a query to the full e-mail collection can be done in real time on a standard PC.

Bursts might pick up

- Trends in language use, rather than content.
- Changes in the construction of text—e.g., in presidential speeches.

Acknowledgments

We would like to thank Miguel Lara and his colleagues at the Center for Innovative Teaching and Learning, University Information Technology Services at Indiana University, Bloomington.

The tool development work is supported in part by the Cyberinfrastructure for Network Science Center (http://cns.iu.edu) and Indiana University, the National Science Foundation under Grants No. SBE-0738111 and IIS-0513650, the US Department of Agriculture, the National Institutes of Health, and the James S. McDonnell Foundation.

Many visualizations used in the course come from the *Places & Spaces: Mapping Science* exhibit, online at http://scimaps.org, and from the *Atlas of Science: Visualizing What We Know*, MIT Press (2010).

INDIANA UNIVERSITY NSF USDA NATIONAL INSTITUTES OF HEALTH James S. McDonnell Foundation CNS Cyberinfrastructure for Network Science Center