Topic Dynamics: An Alternative Model of 'Bursts' in Streams of Topics *

Dan He
UCLA Computer Science Dept.
Los Angeles CA 90095-1596
danhe@cs.ucla.edu

D. Stott Parker
UCLA Computer Science Dept.
Los Angeles CA 90095-1596
stott@cs.ucla.edu

ABSTRACT

For some time there has been increasing interest in the problem of monitoring the occurrence of topics in a stream of events, such as a stream of news articles. This has led to different models of *bursts* in these streams, i.e., periods of elevated occurrence of events. Today there are several burst definitions and detection algorithms, and their differences can produce very different results in topic streams. These definitions also share a fundamental problem: they define bursts in terms of an arrival rate. This approach is limiting; other stream dimensions can matter.

We reconsider the idea of bursts from the standpoint of a simple kind of physics. Instead of focusing on arrival rates, we reconstruct bursts as a dynamic phenomenon, using kinetics concepts from physics — mass and velocity — and derive momentum, acceleration, and force from these. We refer to the result as *topic dynamics*, permitting a hierarchical, expressive model of bursts as intervals of increasing momentum. As a sample application, we present a topic dynamics model for the large PubMed/MEDLINE database of biomedical publications, using the MeSH (Medical Subject Heading) topic hierarchy. We show our model is able to detect bursts for MeSH terms accurately as well as efficiently.

Categories and Subject Descriptors

I.6.5 [**Computing Methodologies**]: SIMULATION AND MODELING—*Model Development, Modeling methodologies*; J.3 [**Computer Applications**]: LIFE AND MEDICAL SCIENCES—*Medical information systems*

General Terms

Theory

*Supported by NIH grants RL1LM009833, UL1DE019580 (UL1RR024911), P20RR020750, P20MH065166, and CCB (UCLA Center for Computational Biology, RO1MH082795).

KDD'10, July 25–28, 2010, Washington, DC, USA.

Keywords

Burst, Hierarchy, PubMed, Momentum, Topic dynamic

1. INTRODUCTION

With the ever-increasing volume of technical publications it is increasingly challenging to stay up-to-date with trends in many fields. Furthermore hot topics or shifts in trends are often manifested in 'bursts' of publications, and these are difficult to stay on top of. It is natural to hope that we might be able to better manage some parts of this continuing challenge by automating the identification of new trends, and the detection and tracking of topic bursts. There has been a great deal of research into methods for identifying and tracking topics 'on the fly', such as topic detection and tracking (TDT) [**?**, **?**, **?**, **?**], modeling the evolution of topic hierarchies over time [**?**], and automatically adjusting (segmenting) time partitions to find temporal contexts [**?**].

In this paper we are concerned with the spread of topics over time through scientific publications. Specifically we will focus on the curated, slowly-evolving hierarchy of *MeSH terms*: a hierarchy of topic keywords integrated with the bioscientific literature indexed by PubMed/MEDLINE (www.nlm.nih.gov/mesh, see Section 4.1). Specifically, we are interested in formalizing the notion of a 'burst' that occurs in a *topic stream* (a stream of topics), and characterizing topic streams in the scientific literature.

In the groundbreaking paper [**?**], Kleinberg developed a framework for doing this, defining a word burst as an increase in the occurrence (arrival rate) of the word in a stream of text, and developing an automaton for tracking an optimized estimate of this rate. His subsequent paper on tracking 'meme' topics in the news [**?**] (`www.meme-tracker.org`) formulates meme topics as patterns of words, and builds on this model of bursts. Kleinberg's burst model has inspired many subsequent efforts, since his traffic-based burst definition is intuitively appealing. Still, this definition may not always be the most appropriate. This is discussed in more detail below (Section 2.1), but for example, this burst model can require modification because the definition is relative to a single stream, and the arrival rate in the definition can be hard to characterize.More generally, defining bursts only in terms of an arrival rate — a one-dimensional model — can be limiting, since other dimensions of 'bursts' can matter. For example, for the scientific literature, the arrival rate of events is much slower and more difficult to characterize than that of the the news feeds considered in [?]. Other dimensions of bursts can be important also, such as the different dimensions of 'impact' of the scientific liter-

ature reviewed in the next section. Furthermore scientific topics are identified differently, such as by taxonomies or ontologies of topics; in particular the PubMed/MEDLINE database (`www.pubmed.gov`) is indexed by the MeSH topic ontology (`www.nlm.nih.gov/mesh`). Although scientific topics in this paper are from MeSH, our methods will work with other ontologies. Existing burst models also have limitations in computationally expensive and vague definitions of burst strength etc. These limitations can be serious for detecting bursts in large biomedical literature databases like PubMed, which contains millions of citations. An alternative model addressing these limitations is needed.

The essence of this paper is to view bursts as intervals of increasing 'momentum'. We can allow topics to take different attributes — such as position, velocity, mass — that can have any underlying meaning we like. We can then track changes of sign in velocity, and therefore changes of sign in momentum (mass · velocity). This is a simple and very general notion of burst, and we believe it is useful. The emphasis on momentum is a stock market view of bursts, and thus also a stock market view of topic streams. Technical stock market analysis [**?**] focuses heavily on monitoring momentum, since it can be a natural measure of human sentiment. Increases in momentum are viewed as bursts of sentiment, and swings in momentum (shifts in its sign) as changes in larger trends. In this paper we highlight advantages of this approach for scientific topics. A market perspective can be appropriate if we view science as a marketplace of ideas.

This approach can offer a number of advantages. An immediate advantage is that we can adapt existing machinery for monitoring trends, identifying burst periods, and measuring burst strength. This instantly provides many familiar tools for analyzing and visualizing patterns, backed by decades of experience.

Other advantages of the momentum model lie in the following aspects:

- Allowing different attributes in the definition of bursts provides a much richer model of bursts, which for example can represent notions of burst strength.
- Technical trend monitoring tools are computationally very efficient, operating in an online fashion, unlike some algorithms implementing Kleinberg's model [?].
- The model can not only detect bursts, but also can be used to predict oncoming bursts based on momentum.
- The model integrates hierarchical topic structure in bursts, which allows semantic information among terms to be considered.

Our experiments on PubMed indicates that this model is a natural and practical way of detecting bursts in scientific topic streams, and permits analysis of hierarchical topic structure in bursts.

The paper is organized as follows: Section 2 discusses some limitations of existing burst models, and the advantages of a simplified approach in computing burst strength given the specific domain. Section 3 (particularly 3.3) details our Topic Dynamics burst model, and Section 4 analyzes this model with experiments using MeSH topics for PubMed, as well as a reproduction of Kleinberg's thought-provoking experiment with VLDB and SIGMOD topics — in both cases contrasting the results of the different burst models. We conclude with an assessment of this 'momentum' model of bursts, and with some directions for further work in topic dynamics.

2. RELATED WORK ON TRACKING TOPICS AND 'BURSTS'

There has been a great deal of research into methods for identifying and tracking topics 'on the fly': topic detection and tracking (TDT) [?, ?, ?, ?], modeling the evolution of topic hierarchies over time [?], and automatically adjusting (segmenting) time partitions to find temporal contexts [?]. Börner et al have also studied specific evolution of hot topics [**?**], diffusion of knowledge among research institutions [**?**], and patterns of coauthorship [**?**]. There also has been a considerable amount of work on 'mapping science' [**?**, **?**], and Cokol and Esteban [**?**] have developed a site called *SciTrends* that maps the volume of publications for any given topic, along with events (significant papers), using plots that are reminiscent of stock market charts. Two particularly important models of bursts have been developed by Kleinberg and by Shasha and co-workers, described in the following two sections.

2.1 Kleinberg's Burst Model

Motivated originally by a problem of representing bursts of email messages, Kleinberg's burst algorithm [?] models bursts with an infinite state automaton in which each state represents a message arrival rate (of a Poisson arrival process). The higher the state, the smaller the expected time gap between messages. 'Word bursts' can then be defined as having arrival rates defined by the number of messages containing a particular word. Additionally, jumping from a lower state to a higher state has an associated cost, while the cost to drop down from a higher state to a lower state is 0. Formally, these states are determined by optimization:

> For a time series $\mathbf{z} = \{ z_t \mid t = 0, 1, \ldots, n \}$ of *inter-arrival gaps*, find a state sequence $\mathbf{q} = \{ q_{i_t} \mid t = 0, 1, \ldots, n \}$ minimizing cost function
>
> $$c(\mathbf{q}|\mathbf{z}) = b(\mathbf{q}) \ln((1-p)/p) + \left(\sum_{t=0}^{n} -\ln f_{i_t}(z_t) \right)$$
>
> where p is the probability of a state change, $b(\mathbf{q})$ is the number of state transitions (changes in successive states) in $\mathbf{q}$, and $f_i(z) = \alpha_i e^{-\alpha_i z}$ is the exponential density function for gap values z with arrival rate α_i.

By finding the optimal sequence of states minimizing the cost of transitions and the cost of differences between real arrival rate and the predicted emission rate, a time series of burst strengths is obtained. The complexity of the algorithm for finding this optimal sequence itself can be high, however; optimizing over the state space can be challenging for large-scale analyses or for online applications.

Kleinberg's recent paper on topic tracking in the news [?] formulates 'memes' as patterns of words, using the model of bursts developed earlier in [?]. A major contribution of the paper is to propose scalable clustering approaches for identifying short distinctive phrases traveling intact through on-line text.

2.2 Shasha's Burst Model

Shasha and co-workers have developed several burst definitions [**?**, **?**] based on hierarchies of fixed-length time intervals, motivated originally by a problem of modeling bursts of gamma rays. These wavelet-like hierarchies have intervals of different scale defined by powers of 2, much as Kleinberg's states correspond to arrival rates defined by the powers of a given parameter s. Bursts occur in intervals in which event frequency of occurrence exceeds a given threshold. Formally, bursts are defined in [?] as optimal windows:

> For a time series $\mathbf{x} = \{ x_t \mid t = 0, 1, \ldots, n \}$, given a set of window sizes $\{ w_1, w_2, \ldots, w_m \}$, an aggregate function F and thresholds $f(w_j)$ associated with each window size $(j = 1, 2, \ldots, m)$, find all subsequences of window sizes such that the aggregate function F when applied to the subsequences exceeds their thresholds, i.e., $\forall i \in \{1, ..., n\}$, $\forall j \in \{1, ..., m\}$, $F(\mathbf{x}_{ij}) \gg f(w_j)$ on the window $\mathbf{x}_{ij} = \{ x_t \mid t = i, \ldots, (i + w_j - 1) \}$.

2.3 Limitations of these Definitions

We believe these existing definitions can be improved for modeling topic bursts in the scientific literature. With publication databases like PubMed — databases of millions of documents that span decades — alternative burst definitions appear to be needed. Some specific limitations highlighted by PubMed/MEDLINE include:

- Biomedical publications are released in batches. As a consequence, time becomes discretized into intervals, and a *histogram* — a summary of arrivals — can be more natural than an *arrival rate*. Topic bursts then become time intervals characterized by significant changes in the proportion (relative frequency) of papers that involve the topic.

- With PubMed, the underlying arrival process is not clearly Poisson, an assumption in Kleinberg's model. It may not be memoryless, and can even be close to deterministic; for example the arrival rate of scientific publications very often increases gradually over time. Without some caution this increase can result in bias towards detection of increased burst strength in more recent years. In a publication database such as PubMed, where many thousands of journal streams are interwoven in discrete intervals, arrival rates may fluctuate considerably, and be difficult to formalize.

- Kleinberg's and Shasha's burst models were defined within the context of a single topic. This notion of burst is applicable in certain domains where independence of streams is a valid assumption, such as examining bursts of activity in a particular stock index. However, when the space of topics is unstructured in this way, all information is carried by frequency changes of the individual topic — and one must use these changes to infer any timeline or underlying causes. This may permit detection of associations between bursts for different topics, but this association is likely to be lost in the background of all other bursts. When a hierarchy of topics is available, we can track bursts among related topics as well as how much a topic contributes to its super-topics' bursts.

- Publication topics evolve [?], and are prone to concept drift, which can raise questions about the notion of a topic 'burst' over longer time periods.

- The two models can be computationally expensive, and may not be practical for detecting bursts in large data stream or database — such as the PubMed biomedical publication database, which now contains more than 19 million citations.

3. TOPIC DYNAMICS: RECONSIDERING BURSTS IN TERMS OF MOMENTUM

In this section we propose an alternative definition of bursts involving 'momentum'. Specifically, a burst is a time interval over which the rate of change of momentum is positive. This rate of change has a natural interpretation as a kind of 'acceleration' or 'force', leading us to an intuitive physical model of bursts.

Instead of defining bursts in terms of arrival rate in a single topic stream, our topic dynamic model adapts basic notions of dynamics from physics and models topic bursts as momentum change of the topic. As defined in physics, momentum is the product of mass and velocity, which is the rate of position change. This model defines mass as the current importance of the topic and position as its intensity (which could reflect arrival rate). Since it is hard to measure the change of momentum from these values directly, we adapt popular stock market trend analysis techniques such as *EMA* (Exponential Moving Average) and *MACD* (Moving Average Convergence/Divergence) in our model. These yield established measures of momentum and popular indicators of trends in dynamic marketplaces (including marketplaces of ideas). Since all the trend analysis indicators are computable in an online fashion, our model can be very efficient, avoiding implementation complexity of existing burst models. The topic dynamic model also leads to a clear definition of burst strength as the *MACD* histogram value. This benefits not only detection of bursts, but also momentum-based prediction of bursts. A hierarchical topic structure is naturally integrated into the model, so that bursts are accumulated along the hierarchical structure. This allows exploration of semantic information involving multiple topics. We show the detailed description and analysis of our topic dynamics model in the following sections.

3.1 Useful Measures of a Topic

One problem with 'bursts' in general is that they can be viewed in very different ways, and formal definitions can rest on vague intuition. We try to reconstruct basic ideas about bursts using intuition about 'momentum'; an interval in which momentum is increasing is called a 'burst'.

A topic in the scientific literature can have many useful quantitative statistics. As the most basic constructs here, we assume that each topic appearing in the stream has two associated quantities:

- a *position* $x(t)$, a linearly-ordered measure or intensity at time t. Examples include: numbers of articles containing the topic, number of pages including the topic, numbers of accesses or downloads, and numbers of pages. This can be a measure of attention devoted to the topic, like arrival rate or communication volume.

- a *mass* $m(t)$, aggregate weight or importance at time t. Examples include: numbers of article citations, journal impact factors, and journal relevance measures. This value can incorporate statistical concepts like variance, but it can be any measure of importance of the topic or of its communication channels.

In general $m(t)$ and $x(t)$ may be defined only at discrete points in time, and (as will be shown shortly) aggregations of $m(t)$ and $x(t)$ may use weightings that emphasize recent points in time, but these provide two dimensions for characterizing bursts. Constant weights (even $m(t) = 1$) are often enough, but $m(t)$ can also be a function of time.

A *velocity* $v(t)$ may be obtainable simply by computing $dx(t)/dt$ in general, but here we have discrete events and thus will estimate $v(t) \simeq \Delta x(t)/\Delta t$. Taking this a step further, we can adapt basic notions of dynamics (kinetics, Newtonian mechanics) from physics to fit with these two basic constructs:

- **mass**: $m(t)$
- **position**: $x(t)$
- **velocity**: $v(t) = dx(t)/dt$
- **momentum**: mass $\cdot$ velocity $\simeq m(t) \cdot v(t)$
- **acceleration**: $dv(t)/dt$
- **force**: mass $\cdot$ acceleration $\simeq m(t) \cdot dv(t)/dt$

3.2 Momentum as 'Impact'

Of these measures, momentum is perhaps the most basic – $m(t) \cdot v(t)$ or equivalently the rate of change of the product $x(t) \cdot m(t)$ — a measure of change in 'impact'. In any marketplace of ideas (whether the stock market or the scientific literature), change in impact is important. For stocks, $x(t)$ can be a measure of 'value' like stock price or P/E ratio. and $m(t)$ can be a measure of market 'importance', such as trading volume or market capitalization (share price times number of outstanding shares). This market perspective can work well for topics, especially when 'total communication volume' of the topic matches the idea of its 'impact'. This perspective has yielded interesting results in the past for the scientific literature. For example, Cokol and Esteban [?] built a system called SciTrends (`www.scitrends.com`) that permits visualization of science publication trends using charts that are highly reminiscent of the stock market.

With the physical notions just defined, we can model a *burst* in terms of momentum, (or equivalently, if not weighted, in terms of intervals of positive *acceleration*). Specifically, a burst is a time interval over which acceleration is positive. If we are looking at a value like *position* $x(t)$, in other words, a burst is an interval where $v(t) = dx(t)/dt$ is increasing, or equivalently, $dv(t)/dt$ is positive. Thus, a burst is a period of positive acceleration; and if weighted by mass, a burst is an interval of positive force.

This leads us to a general model of *topic dynamics*. As a result, momentum is strictly increasing over any burst interval. We define a *burst* as a time interval of maximal length over which the rate of change of momentum is positive. These definitions can go beyond measuring the length and strength of the increased frequency of a particular event, word or topic in a single independent stream. The discrete time values used here and noise in the values of $x(t)$ can make it hard to define quantities like $dx(t)/dt$. In this case we need to use some method for making the derivatives meaningful. The *MACD* indicator described in the next section give a classical means for finding 'trends' in noisy time series, and specifically trends related to momentum and acceleration.

3.3 Formalizing Bursts with Momentum Indicators

We first review some basic estimators of momentum drawn from technical analysis of stocks. Certainly care is needed in adapting tools for stock market analysis to the analysis of momentum in scientific topics. However, there are many sophisticated tools for estimating momentum.

In the stock market, because of the noisiness of values of $x(t)$ like 'price', some smoothing mechanism is often used to permit analysis. A traditional scheme for identifying trends in stock market analysis is to use *moving averages* to smooth out noise, and then estimate derivatives (rates of change, hence trends) using differences of smoothed values (moving averages) that have been computed in mildly different ways. We reproduce basic concepts here (for more see e.g., [?]).

- **EMA**: For a variable $x = x(t)$ which has a corresponding discrete time series $\mathbf{x} = \{ x_t \mid t = 0, 1, \ldots, \}$, the n-day *EMA (Exponential Moving Average)* with smoothing factor α:

$$\begin{aligned} EMA[x]_t &= \alpha\, x_t + (1-\alpha)\, EMA[x]_{t-1} \\ &= \sum_{k \geq 0} \alpha\, (1-\alpha)^k\, x_{t-k}. \end{aligned}$$

  Often this sum is terminated after $n > 0$ terms, in which case n is called the *window size*:

$$\begin{aligned} EMA(n)[x]_t &= \alpha\, x_t + (1-\alpha)\, EMA(n-1)[x]_{t-1} \\ &= \sum_{k=0}^{n} \alpha\, (1-\alpha)^k\, x_{t-k}. \end{aligned}$$

  In this case the factor α is often taken to be $\alpha_n = 2/(n+1)$.

  Usually we are interested in averages of a single indicator value x up to the current time t, so by convention we omit both $[x]$ and t, so for example $EMA(n)$ is a shorthand for $EMA(n)[x]_t$, denoting an *EMA* up to the current time.

- **MACD**: In technical stock market analysis [?], the *MACD (Moving Average Convergence/Divergence)* of a variable x_t (usually price) is defined by the difference of its n_1- and n_2-day moving averages:

$$MACD(n_1, n_2) = EMA(n_1) - EMA(n_2)$$

  This difference is a popular estimate of $\Delta x/\Delta t$, and hence (using our physics terminology) an estimate of velocity. Notice that this difference changes sign when the plots of the two moving averages cross, as shown in Figure 1.

- **MACD histogram**: *MACD histogram* is an estimate of the derivative of the *MACD* :

$$\begin{aligned} \text{signal}(n_1, n_2, n_3) &= EMA(n_3)[MACD(n_1, n_2)] \\ \text{histogram}(n_1, n_2, n_3) &= MACD(n_1, n_2) - \text{signal}(n_1, n_2, n_3) \end{aligned}$$

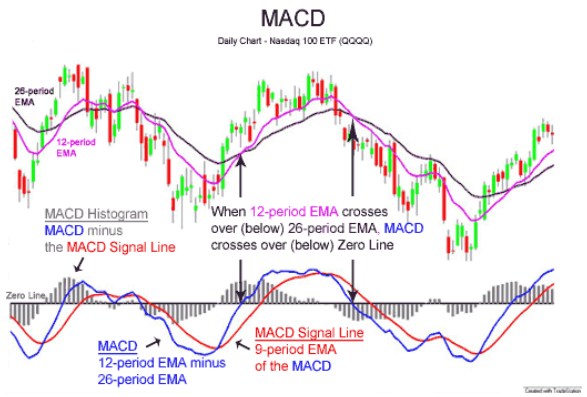

Figure 1: Typical *MACD* plot, showing also both the *MACD* signal line and *MACD* histogram (12,26,9); from en.wikipedia.org/wiki/File:MACDpicwiki.gif.

$EMA(n_3)[MACD(n_1, n_2)]$ denotes the n_3-day moving average of the sequence $MACD(n_1, n_2)$. The histogram is a difference between the *MACD* and its moving average, and so is a kind of 'second derivative'. This difference is interpreted as a measure of change in momentum, and can be both positive and negative. The sign of the histogram changes when the *MACD* value crosses over the signal value — a 'crossover'. In technical stock market analysis, crossovers are interpreted as changes in trends and thus possible encouragements for trades. The difference between these two terms reflects change in the direction of price activity, and the signal is an averaged measure of change. The *MACD* histogram in fact amplifies the difference between the average change and local change. Thus, the histogram measures the rate of change in direction of price activity, i.e., change in momentum.

3.4 Linearity of *MACD* and Momentum

Weighted versions of the *EMA* and *MACD* indicators are common, replacing the values α by an explicit set of weights:

$$\begin{aligned} WEMA_n[w,x]_t &= \sum_{k=0}^{n-1} \alpha_n \; w_k \; x_{t-k} \\ WMACD(p,q)[w,x]_t &= WEMA_p[w,x]_t \; - \; WEMA_q[w,x]_t \end{aligned}$$

For example, the weights w can be the mass values $m(t)$, and the weighted signal is momentum relative to these. In our physical terminology, the *MACD* histogram measures *acceleration* or *force*. As momentum is measured with velocity $v(t)$ (or mass times velocity, $m(t) \cdot v(t)$), changes in momentum are measured with acceleration $a(t) = dv(t)/dt$ (or mass times acceleration $m(t) \cdot a(t)$). The *MACD* histogram is therefore a measure of acceleration, and the *WMACD* histogram a measure of force. Zeroes of the histogram demarcate changes in the sign of momentum, and changes in trends.

PROP 1. *The MACD, the MACD signal, the MACD histogram and their weighted counterparts are all* linear functions *of a time series, in the sense that each maps a time series* $\boldsymbol{x} = \{ x_t \mid t = 0, 1, \ldots, \}$ *to a time series* $\boldsymbol{y} = f(\boldsymbol{x})$ *defined by a linear combination*

$$y_t \;\; = \;\; \sum_{k=0}^{n} w_k \; x_{t-k}.$$

This linearity follows from the fact that the *EMA* is linear in this sense, and the difference of two such linear combinations is also a linear combination.

For example, if $\alpha_n = 2/(n+1)$ as indicated above,

$$\begin{aligned} &MACD\ histogram(4,8,5) \\ &= \;\; (1-\alpha_5)\,(\alpha_4 - \alpha_8) \;\; x_t \\ &+ \;\; (1-\alpha_5)\,(\alpha_4 - \alpha_8)\,(1 - \alpha_5 - \alpha_4 - \alpha_8) \;\; x_{t-1} \\ &+ \;\; \ldots \\ &+ \;\; \alpha_5 \alpha_8\,(1-\alpha_5)^4\,(1-\alpha_8)^7 \;\; x_{t-11} \\ \\ &= \; \tfrac{16}{135}\, x_t + \tfrac{32}{6075}\, x_{t-1} - \tfrac{9536}{273375}\, x_{t-2} + \cdots + \tfrac{26353376}{10460353203}\, x_{t-11}. \end{aligned}$$

This formula highlights several important aspects of our burst model. At a fundamental level, we are defining bursts as a kind of *linear filter*. Beyond the usefulness of technical analysis for identifying trends and bursts in noisy market data, defining bursts in terms of a family of well-understood pattern detectors (linear filters) can bring many analytical tools for data mining.

Furthermore, our burst model is parametric, and the parameters can be tuned to the application. We do not discuss the problem of learning here, but we can train our burst model to match a set of burst examples with an optimized search through the *EMA*/*MACD* parameter space. It can be advantageous to have a vector space of burst models defined by parameters of this kind, since it permits definition of bursts in a problem-specific fashion.

PROP 2. *Given two MACD histogram models* (p,q,r) *and* (s,t,u)*, the symbolic form of their difference* $\sum_{k=0}^{n} d_k \; x_{t-k}$ *specifies intervals on which their burst models differ: sequences* $\boldsymbol{x}$ *for which the difference is nonzero define patterns for which the two models yield different burst windows.*

That is, we can compare burst models qualitatively by analyzing their linear form. For example, in some situations we have used a $(7,9,5)$ instead of a $(4,8,5)$ *MACD* histogram. Consider their difference for a sequence $\mathbf{x}$:

$$\begin{aligned} &MACD\ \text{histogram}(7,9,5) \;\; - \;\; MACD\ \text{histogram}(4,8,5) \;\; = \\ &\underline{-0.0852}\, x_t \;\; + \;\; 0.00195\, x_{t-1} \;\; + \;\; 0.0298\, x_{t-2} \;\; + \;\; 0.0320\, x_{t-3} \\ &+ \;\; 0.0595\, x_{t-4} \;\; + \;\; 0.0197\, x_{t-5} \;\; + \;\; 0.000904\, x_{t-6} \\ &\underline{-\;0.0292}\, x_{t-7} \;\; \underline{-\;0.0385}\, x_{t-8} \;\; + \;\; 0.000171\, x_{t-9} \\ &+ \;\; \underline{0.00262}\, x_{t-10} \;\; + \;\; \underline{0.00356}\, x_{t-11} \;\; + \;\; 0.00221\, x_{t-12}. \end{aligned}$$

The difference has only 3 negative terms, and these put particular emphasis on the most recent value x_t. Thus the $(7,9,5)$ histogram value is generally higher than the $(4,8,5)$ value unless x_t is relatively large, and thus it will obtain larger burst windows. Generally the $(7,9,5)$ histogram profile will be smoother, and emphasize history, while $(4,8,5)$ will emphasize the most recent value.

Finally, momentum of a time series is naturally a linear function, involving weighted averages (with $m(t)$) and differences (with $dx(t)/dt$). In this paper we use *MACD* indicators, but other linear functions could be used. There are many kinds of moving average, and many different numerical approximations for derivatives. The *MACD* indicators are well-established tools of technical analysis, and have worked well for us in identifying bursts.

```
C01.539               Infection
C01.539.778               Sexually Transmitted Diseases
C01.539.778.281               Sexually Transmitted Diseases, Bacterial
C01.539.778.281.201               Chancroid
C01.539.778.281.301               Chlamydia Infections
C01.539.778.281.401               Gonorrhea
C01.539.778.281.451               Granuloma Inguinale
C01.539.778.281.859               Syphilis

C02          Virus Diseases
C02.081          Arbovirus Infections
C02.109          Bronchiolitis, Viral
C02.182          Central Nervous System Viral Diseases
C02.256          DNA Virus Infections
C02.290          Encephalitis, Viral
C02.325          Eye Infections, Viral
C02.330          Fatigue Syndrome, Chronic
C02.407          Hepatitis, Viral, Animal
C02.440          Hepatitis, Viral, Human
C02.587          Meningitis, Viral
C02.597          Opportunistic Infections
C02.705          Pneumonia, Viral
C02.782          RNA Virus Infections
C02.800          Sexually Transmitted Diseases
C02.800.801          Sexually Transmitted Diseases, Viral
C02.825          Skin Diseases, Viral
C02.839          Slow Virus Diseases
C02.928          Tumor Virus Infections
C02.937          Viremia
C02.968          Zoonoses
```

hierarchy depth	number of MeSH terms
1	109
2	1495
3	6527
4	12446
5	12702
6	7960
7	4304
8	1820
9	800
10	242
11	38

Figure 2: **Sample segments of the MeSH term hierarchy, including terms above and below 'Sexually Transmitted Diseases'. The histogram gives the number of terms at each level of the 2008 version of the MeSH hierarchy of topics, comprising altogether 48,443 terms. This hierarchy is used as a curated keyword indexing mechanism for PubMed/MEDLINE.**

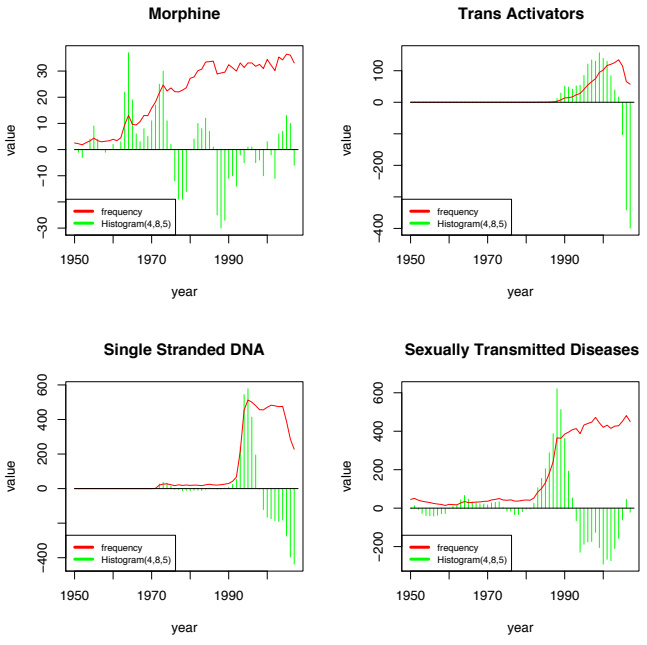

Figure 3: **Frequency and *MACD* histogram(4,8,5) for the terms 'Morphine', 'Trans Activators', 'Single Stranded DNA', 'Sexually Transmitted Diseases'.**

4. ANALYZING PUBMED TOPIC BURSTS

We analyzed our topic dynamic model using PubMed, a biomedical publication database that includes over 19 million citations, primarily from life science journals. We developed our model using the MeSH hierarchy, which contains more than 50,000 terms arranged in a generalization hierarchy. We first show the properties of our model on a subhierarchy of MeSH terms, then we analyze the burst detection accuracy of our model on specific terms against historical events to show our model is able to detect bursts accurately. We also compare our method with Kleinberg's burst detection method and we show our method can be more practical for topic bursts detection in biomedical literature.

4.1 The MeSH Hierarchy

MeSH (Medical Subject Headings) is a hierarchy of topics integrated with PubMed/MEDLINE (`www.nlm.nih.gov/mesh`). The 2008 version of MeSH contains about 50,000 terms arranged in a generalization hierarchy with the depth histogram as shown in Figure 2. Notice that this term occurs twice in the hierarchy, with parents 'Infection' and 'Virus Diseases' — and different children depending on the context. Thus MeSH is not a tree; it is a directed acyclic graph.

For scientific publications, a natural window width is 1 year. We consider both finer and coarser resolutions below, but they can create problems, including artifactual bursts and over-smoothing. We therefore have aggregated term frequency counts over annual time windows. From each article citation in PubMed, we extract n tuples $T = (y, m)$ where y is the published year of the citation, m is the MeSH term, and n is the number of MeSH terms associated with the citation. For each MeSH term m, we look up its set A of ancestor MeSH terms to produce $|A|$ new tuples. Next, from all tuples generated from this citation, duplicates are

removed using a hash table. Then, for each tuple $T = (y, m)$ generated, the frequency count is incremented by 1.

4.2 Burst Analysis for MeSH

We applied our model on all 50,000 terms in the 19 million biomedical publication citations. Since our model uses online computation, it finishes in a few minutes. To show our model makes reasonable and accurate burst detections, we checked the detections of bursts for certain terms which are related to known important events. For illustration purpose, frequency is normalized with a scale factor to fit into the value range of histogram.

The rule parser flagged the topic 'Morphine' during 1965-1975 as a bursty period in which no child topics were bursty. It is interesting to note that the Vietnam War started in 1963 and ended in 1975. One of the periods of interesting burstiness in Figure 3 corresponds almost exactly to the Vietnam War. From the frequency plot in Figure 3, we see that towards the end of the war, publications on morphine slowed, resulting in a drop of burst strength. (Since frequency counts for 2008 were incomplete, there is also a sharp drop in both frequency and histogram values in 2008.)

During the early 1990s, a flurry of genetic advancements made deeper research on DNA possible. From the start of the Human Genome Project in 1990 to the cloning of Dolly the sheep in 1996, the growth of genetic research generated bursts in DNA-related topics during this period. Trans-Activators and Single-Stranded DNA are two DNA-related topics that were detected by the parent-child-context rule. Trans-activators are diffusible gene products that act on molecules of viral or cellular DNA to regulate the expression of proteins, according to the *Online Medical Dictionary* (`cancerweb.ncl.ac.uk/cgi-bin/omd`). The detected period of interesting burstiness is 1986-2005 with a sharp increase at 1994 as confirmed by Figure 3. This detected bursty time period corresponds with the start of the era of genetic breakthroughs. Single-stranded DNA was detected as an interesting burst during 1988-1998. In fact, we can see there is an increase in burst strength starting at 1988, and the strength increases sharply at 1993 from Figure 3. The burst periods of both terms start before 1990, indicating more attention has been paid to the DNA-related fields, which may lead to the above-mentioned two projects. The success of the two projects boosted people's interest, resulting in sharp increases of publications. Another thing observed is that the trend of the histogram clearly precedes the trend of the frequency for both terms, suggesting that the histogram can be useful in predicting oncoming events.

As shown in Figure 3, the period of dramatic increase in burst strength corresponds to the period of increased awareness regarding Sexually Transmitted Diseases. For instance, in 1986-1987, the FDA approved the first anti-retroviral drug to treat AIDS [**?**]. Hence, real world events are consistent with changes in burst strength starting in 1982. The burst reaches its peak after FDA approval.

We further verify our burst detection model with the terms identified as bursty words in [**?**]. In this work, the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982 to 2001 were identified. Their most significant burst periods were also listed. We select four terms 'Signal Transduction', 'Nucleic Acids', 'Antibodies' and 'Molecular Sequence Data' from these top 50 words such that their burst periods are all different from each other. We apply our burst detection model and compare our detection with their burst periods identified in [?]. The results are shown in Figure **??**. As we can see, the detections of our model are all consistent with these pre-identified burst periods. What's more, the peaks of the bursts detected by our model all fall in the pre-identified burst periods. This again confirms that using the *MACD* histogram can be effective for identifying bursts.

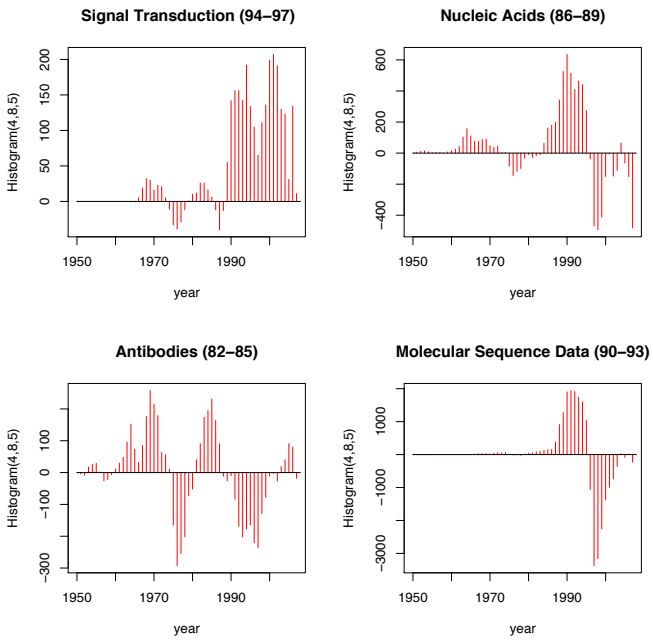

Figure 4: *MACD* **histogram(4,8,5) for the terms 'Signal Transduction', 'Nucleic Acids', 'Antibodies' and 'Molecular Sequence Data'. Their bursty periods as shown in [?] are 1994-1997, 1986-1989, 1982-1985 and 1990-1993, respectively.**

4.3 Comparison with Kleinberg's burst model

Kleinberg's burst algorithm was reportedly used in [?], in studying bursts of words appearing in articles related to Melanoma. Specifically, it was used to detect bursts of names for genes and proteins using PubMed. We performed a similar analysis for these words using PubMed, and the results are shown in Figure **??**. If our method of gathering data is in fact the same as that in [?], then there are significant differences between the 'bursts' detected by the two different methods. Furthermore, the time series of frequencies of the four gene names shown in Figure **??** seem inconsistent with the bursts detected using Kleinberg's algorithm. Admittedly, this algorithm can use much more precise timing information than the annual sampling we used, and the PubMed data gathered for [?] in 2004 may have changed somewhat. Nevertheless, it appears the intervals detected using momentum are different, and we believe that they better reflect the structure of the word frequency time series.

Kleinberg's model associates an automaton characterized by the arrival rate for each term and identifies bursts for terms with higher arrival rate than those of others in the

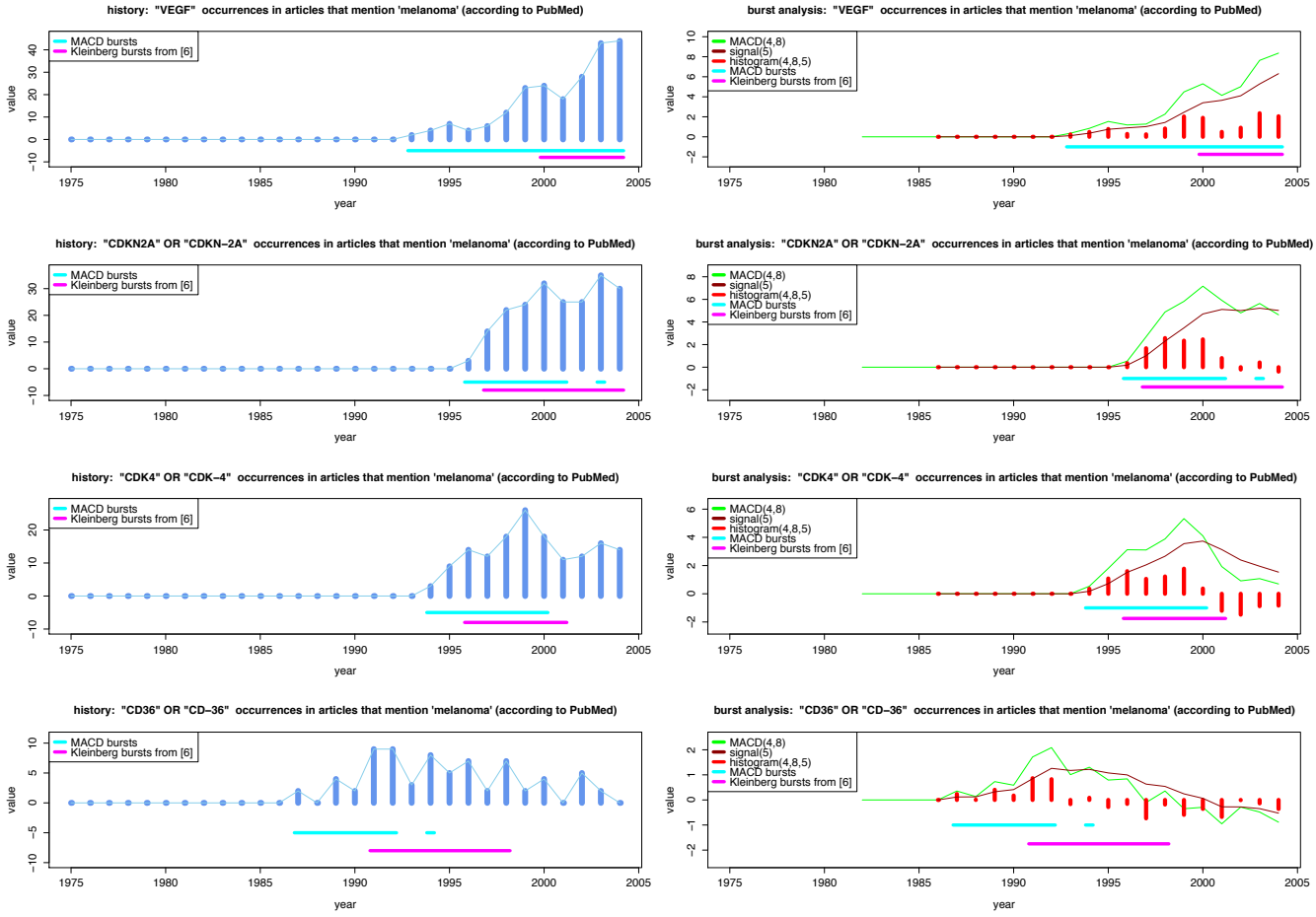

Figure 5: Analysis of some of the gene names described as having Kleinberg-style bursts in Figure 2 of [?], showing on the left the time series of frequencies (annual occurrence histogram) of each gene name (PubMed query) and on the right the corresponding burst analysis using momentum. Cyan intervals at the base of vertical histogram bars identify years in a MACD burst, while magenta intervals identify the Kleinberg bursts reported in [?]. The burst intervals differ significantly, raising questions about the nature of the two different kinds of 'bursts'. Here a MACD burst is found in 1993–2004 for VEGF (vs. a Kleinberg burst in 2000–2004), 1996–2001 for CDKN2A (vs. 1997–2004), 1994–2000 for CDK4 (vs. 1996–2001), and 1987–1992 for CD36 (vs. 1991–1998). Notice also that the histogram(4,8,5) on the right also captures burst strength.

same stream of text. Thus the model is stream-dependent. Our model, on the contrary, is stream-independent, in that it only utilizes the historical arrival rate of each term and identifies significant changes in the arrival rate as bursts. To illustrate such difference, we ran our model on the terms in the paper titles from the database conferences SIGMOD and VLDB based on the frequencies of the terms, for the years 1975-2001, in the same manner as Kleinberg [?]. Using histogram(4,8,5), we compared the burst intervals for the terms of highest weight from Kleinberg's model with those from our model (due to space restriction, we select only 12 terms). The comparison is shown in Figure 6. Notice that we show only the highest weight burst intervals from Kleinberg's model but show all burst intervals of these terms from our model. Consequently there are significantly more burst intervals from our model in the table. However, by comparing the intervals over similar years, we observe that the burst intervals from Kleinberg's model are usually long, spanning multiple years, while the burst intervals from our model are relatively short. During these intervals, the arrival rates of these terms are high while the arrival rates of other terms are relatively low. Thus in Kleinberg's model once the arrival rates of 'bursty' terms become relatively high their burst intervals continue, whereas in our model (since it doesn't compare arrival rates between terms) any significant increase detected by the histogram in the arrival rate of any term is reported as a burst.

To summarize, Kleinberg's model identifies bursts in a stream of text mixed with multiple terms, while our model identifies bursts in a stream of only single terms. The two models address different problems. In our MeSH hierarchy setting, we aggregate the frequencies of child topics with their ancestors. Therefore, the ancestors always have much higher arrival rates and thus tend to have bursts according to Kleinberg's model, as it identifies terms with higher arrival rates relative to other terms. Thus we believe our term-independent model is more appropriate for burst detection in the biomedical literature using MeSH terms.

word	Bursts (Kleinberg)	Bursts (Our model. S = *SIGMOD*, V = *VLDB*)
bases	$1975SIGMOD - 1982VLDB$	$1975V, 1977V, 1978V, 1979V, 1980V, 1982V, 1984V - 1989S, 1990S - 1991V, 1993V, 1995V, 1998S$
object	$1990SIGMOD - 1996VLDB$	$1983V - 1984V, 1986S - 1987V, 1989S - 1992S, 1994S, 2000S$
parallel	$1989VLDB - 1996VLDB$	$1979V - 1980S, 1982V, 1986V - 1987S, 1988S, 1989S - 1990V, 1991V - 1992S, 1993S - 1994S, 1995V - 1996V, 2001V$
statistical	$1981VLDB - 1984VLDB$	$1977S - 1977V, 1979V, 1981V - 1982V, 1987V - 1990V, 1994S - 1995V, 1997V, 2001V$
model	$1975SIGMOD - 1992VLDB$	$1975V - 1976S, 1978V - 1979S, 1981V, 1982V, 1983V, 1985S - 1985V, 1988S, 1989S, 1990V, 1991V, 1992V, 1995V, 1996V, 1998S$
schema	$1975VLDB - 1980VLDB$	$1975V - 1978V, 1983V, 1984V - 1985V, 1987S - 1987V, 1988V, 1989V, 1991V, 1993V, 1994V, 1995V, 1998S - 1999S, 2000V$
web	$1998SIGMOD-$	$1995S - 1996S, 1997S - 1998S, 1999S$
approximate	$1997VLDB-$	$1982S - 1982V, 1986V - 1994S, 1997V - 1999S, 2000V$
objects	$1987VLDB - 1992SIGMOD$	$1984V - 1985V, 1986V - 1988S, 1989S - 1989V, 1991S, 1995S, 1996S, 1997V, 1998V$
xml	$1999VLDB-$	$1999V$
warehouse	$1996VLDB-$	$1994S - 1995V, 1996V - 1997S, 1999S, 2000S$
server	$1996SIGMOD - 2000VLDB$	$1984V - 1985S, 1987V, 1991V - 1993S, 1994S - 1995S, 1996S - 1996V, 1997V - 1998V, 1999V$

Figure 6: Comparison between Kleinberg's burst intervals [?] and our burst intervals on the paper titles from the database conferences SIGMOD and VLDB.

word	*parameters*	Burst Intervals (Our model. S = *SIGMOD*, V = *VLDB*)
database	(4, 8, 5)	1975V - 1976V, 1977V - 1979S, 1980S ,1981S ,1982V ,1984S ,1985S - 1985V, 1987S - 1988V, 1991S ,1992S - 1995S, 1996S ,1998V ,2001S
	(7, 9, 6)	1975V - 1981S, 1985S - 1985V, 1987S - 1988V, 1992S - 1995S, 1996S ,2001S
transaction	(4, 8, 5)	1975V - 1976S, 1979S - 1979V, 1981V - 1982S, 1984V - 1985S, 1987S - 1988S, 1990V ,1991V - 1992S, 1995V ,1997V - 1998S, 2000V
	(7, 9, 6)	1975V - 1976S, 1979S - 1980S, 1981V - 1982V, 1984V - 1985S, 1987S - 1988V, 1990V - 1992S, 1995V, 2000V
parallel	(4, 8, 5)	1979V - 1980S, 1982V ,1986V - 1987S, 1988S ,1989S - 1990V, 1991V - 1992S, 1993S - 1994S, 1995V - 1996V, 2001V
	(7, 9, 6)	1979V - 1980S, 1982V - 1983S, 1986V - 1988S, 1989S - 1990V, 1991V - 1994S, 1995V

Figure 7: Comparison between the burst intervals from our model using different histogram parameters on the paper titles from the database conferences SIGMOD and VLDB.

As we discussed before, our burst model has another advantage in that it is parametric, and the parameters can be tuned to the application. To illustrate how different parameters affects the results, we conducted the same experiments on paper titles from SIGMOD and VLDB, but with different histogram parameters. As analyzed in Section 3.5, the burst intervals using histogram(7,9,6) are generally larger as well as smoother than that of histogram(4,8,5). We show the experiments in Figure 7 (due to space restriction, we only show the comparison for three terms). Obviously histogram(7,9,6) reports larger burst intervals.

5. CONCLUSION

We presented a Topic Dynamics framework as an alternative to detection and analysis of bursts. This framework rests on physical intuition, modeling bursts as intervals of increasing momentum, which can be applied to many 'trend' quantities of interest, such as changes of 'value' in the stock market and changes of 'impact' in the scientific literature.

Our experiments show the model works well for tracking topic bursts of MeSH terms in the bioscientific literature, where arrival rates are often non-random. In addition, the momentum-based topic dynamics burst model described here has significant advantages: (1) momentum can be monitored online using technical stock market analysis tools like the *MACD*, and bursts can be detected using *MACD* histogram, giving a way to adapt the vast domain of technical analysis to burst analysis; (2) the topic dynamics model provides two dimensions of modeling (position $x(t)$ and mass $m(t)$), which can be useful in defining burst strength; (3) the *MACD* formalism permits burst patterns to be represented as linear filters, permitting both qualitative and quantitative analysis of burst patterns.

The topic dynamics framework also has other interesting aspects; it not only detects burst periods, and burst strength, but also can be used for forecasting oncoming bursts (just as momentum is used for forecasting in the stock market). Furthermore, this framework embraces hierarchical structure of bursts, and takes into account semantic links between topics that are missed by single-steam analysis. Elsewhere we have analyzed hierarchical structure in bursts, which fits our model well since momentum is naturally hierarchical. We believe hierarchical structure deserves greater attention in future work on bursts, as it is a central aspect of existing frameworks for burst analysis.

6. REFERENCES

[1] James Allan. *Introduction to topic detection and tracking*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[2] Katy Börner, Luca Dall'Asta, Weimao Ke, and Alessandro Vespignani. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, 10(4):57–67, 2005.

[3] Katy Börner, Shashikant Penumarthy, Mark Meiss, and Weimao Ke. Mapping the diffusion of scholarly knowledge among major U.S. research institutions. *Scientometrics*, 68(3):415–426, 2006.

[4] Kevin W. Boyack, Richard Klavans, and Katy Börner. Mapping the backbone of science. *Scientometrics*, 64(3):351–374, 2005.

[5] Kevin W. Boyack, Ketan Mane, and Katy Börner. Mapping medline papers, genes, and proteins related to melanoma research. In *IV*, pages 965–971. IEEE Computer Society, 2004.

[6] Murat Cokol and Raul Rodriguez-Esteban. Visualizing evolution and impact of biomedical fields. *Journal of Biomedical Informatics*, 41:1050–1052, 2008.

[7] J. Kleinberg J. Leskovec, L. Backstrom. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July, 2009, Paris, France*, 2009.

[8] MI Johnston and DF Hoth. Present status and future prospects for HIV therapies. *Science*, 260(5112):1286–1293, 1993.

[9] Jon M. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada, pages 91–101. ACM, 2002.

[10] Ketan K. Mane and Katy Börner. Mapping topics and topic bursts in pnas. *PNAS*, 101:5287–5290, 2004.

[11] Taneli Mielikäinen, Evimaria Terzi, and Panayiotis Tsaparas. Aggregating time partitions. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 347–356, New York, NY, USA, 2006. ACM.

[12] Satoshi Morinaga and Kenji Yamanishi. Tracking dynamics of topic trends using a finite mixture model. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 811–816, New York, NY, USA, 2004. ACM.

[13] John Murphy. *Technical Analysis of the Financial Markets*. Prentice Hall, 1999.

[14] Colin Murray, Weimao Ke, and Katy Börner. Mapping scientific disciplines and author expertise based on personal bibliography files. In *IV*, pages 258–263. IEEE Computer Society, 2006.

[15] Rene Schult and Myra Spiliopoulou. Expanding the taxonomies of bibliographic archives with persistent long-term themes. In *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing*, pages 627–634, New York, NY, USA, 2006. ACM.

[16] Russell Swan and James Allan. Extracting significant time varying features from text. In *CIKM '99: Proceedings of the eighth international conference on Information and knowledge management*, pages 38–45, New York, NY, USA, 1999. ACM.

[17] Russell Swan and James Allan. Automatic generation of overview timelines. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56, New York, NY, USA, 2000. ACM.

[18] Xin Zhang and Dennis Shasha. Better burst detection. In Ling Liu, Andreas Reuter, Kyu-Young Whang, and Jianjun Zhang, editors, *ICDE*, page 146. IEEE Computer Society, 2006.

[19] Yunyue Zhu and Dennis Shasha. Efficient elastic burst detection in data streams. In Lise Getoor, Ted E. Senator, Pedro Domingos, and Christos Faloutsos, editors, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pages 336–345. ACM, 2003.