Fall 2014

A framework for emerging topic detection in biomedicine

Charisse Renee Madlock-Brown
University of Iowa

A FRAMEWORK FOR EMERGING TOPIC DETECTION IN BIOMEDICINE

by

Charisse Renee Madlock-Brown

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Informatics (Health Informatics)
in the Graduate College of
The University of Iowa

December 2014

Thesis Supervisor: Associate Professor David Eichmann

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Charisse Renee Madlock-Brown

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Informatics (Health Informatics) at the December 2014 graduation.

Thesis Committee: _____
David Eichmann, Thesis Supervisor

_____
James Torner

_____
Christopher Morphew

_____
Gautam Pant

_____
Kang Zhao

To my lovely daughter, Marci.

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 in our past. It does not stretch back evenly, but piles up in 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. Forester
Aspects of the Novel

ACKNOWLEDGMENTS

I would like to thank my family, friends, and advisor for supporting me while I completed this work.

SCIENTIFIC ABSTRACT

Emerging topic detection algorithms have the potential to assist researchers in maintaining awareness of current trends in biomedical fields—a feat not easily achieved with existing methods. Though topic detection algorithms for news cycles exist, several aspects of this particular area make applying them directly to scientific literature problematic.

This dissertation offers a framework for emerging topic detection. It builds upon the probabilistic burst detection algorithm developed by Kleinberg. STC and Lingo Clustering algorithms were used to create an overlapping hierarchical structure of scientific literature at the discipline level. This allows for granularity adjustment (e.g. discipline level or research area level) in emerging topic detection for different users. Using cluster analysis allows for the identification of terms that may not be included in annotated taxonomies, as they are new or not considered as relevant at the time the taxonomy was last updated. Characterization of bursts over an extended planning horizon by discipline was performed to understand what a typical burst trend looks like in this space to better understand how to identify important or emerging trends.

The framework includes a novel set of topic frequency weightings based on the historical importance of each topic identified. The weightings are current term frequency-inverse archive frequency (CTF-IAF), current term frequency -square root-inverse archive frequency (CTF-SQRT-IAF), and for current inverse term frequency-inverse archive frequency (CITF-IAF). These measures were compared to the un-weighted frequency (current term frequency, or CTF). All three weightings identified rare topics as bursty which the CTF measure did not, based on their novelty. CITF performed the best at finding bursty topics which remain bursty, and have the highest future citation rates. Each measure's performance was compared at differing levels of granularity. CITF

performs the best on topics with low frequency counts. At higher frequency counts, CTF-SQRT-IAF, CTF-IAF, and CTF, performed best.

Frequency counts were further weighted with measures such as normalized journal impact factor, normalized h-index, and normalized funding to develop a fitness score to identify which topics are likely to burst in the future. Each fitness score was able to detect bursts earlier for each frequency measure.

PUBLIC ABSTRACT

Emerging topic detection algorithms have the potential to assist researchers in maintaining awareness of current trends in biomedical fields—a feat not easily achieved with existing methods. Though topic detection algorithms for news-cycles exist, several aspects of this particular area make applying them directly to scientific literature problematic.

This dissertation offers a framework for emerging topic detection in biomedicine. The framework includes a novel set of weightings based on the historical importance of each topic identified. Features such as journal impact factor and funding data are used to develop a fitness score to identify which topics are likely to burst in the future. Characterization of bursts over an extended planning horizon by discipline was performed to understand what a typical burst trend looks like in this space to better understand how to identify important or emerging trends. Cluster analysis was used to create an overlapping hierarchical structure of scientific literature at the discipline level. This allows for granularity adjustment (e.g. discipline level or research area level) in emerging topic detection for different users. Using cluster analysis allows for the identification of terms that may not be included in annotated taxonomies, as they are new or not considered as relevant at the time the taxonomy was last updated. Weighting topics by historical frequency allows for better identification of bursts that are associated with less well-known areas, and therefore more surprising. The fitness score allows for the early identification of bursty terms. This framework will benefit policy makers, clinicians and researchers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

Thomas Kuhn's structure of scientific revolutions captures the development of science [1]. According to Kuhn, science progresses by means of revolutionary ideas, which drastically alter the shape of the scientific world. Though many changes occur by means of small incremental developments, there are ideas that create abrupt and dynamic shifts in focus. Large-scale examples of scientific revolutions include the discovery of DNA, and Einstein's theory of relativity. On a smaller scale, discipline, or domain-level revolutions happen all the time from major breakthroughs to minor discoveries [2]. New research topics, which can be a specific area of study, method or tool studied in the scientific community, are constantly altering scientific thought. Maintaining awareness of trends at different levels of granularity is a must for the biomedical community. The question, however, is how this is to be accomplished. My goal is to develop an emerging topic detection framework.

Policy makers, funding agencies and researchers need tools for emerging trend analysis of scientific research. The rate of scientific publication is increasing every year and it is impossible to stay completely up-to-date with traditional methods. NIH program officers need to understand the scope of research fields despite the complexities of the scientific production process. Department and university administrators could more effectively evaluate institutional publication history by determining the relevance of investigators' fields of research. Scientists, who are using online tools to scan more and read less [3] need better methods to stay current. Existing search engines are most useful for aggregation of articles but not detecting new, reemerging, or increasingly popular trends. Being able to detect emerging topics would allow researchers to stay competitive by assisting in finding current related topics to help expand or

support their own work. Clinicians, providing service at point-of-care, will additionally benefit by being able to reevaluate existing patient care based on the newest developments.

Emerging topics are characterized by an increase in topic occurrence frequency in a document stream. Developing successful emerging topic detection methods for scientific work has proved difficult [4]. Though topic detection algorithms for news-cycles exist, several issues make applying them directly to scientific literature problematic. At any given point there are numerous topics in varying stages of the developmental process. At no time does the scientific literature converge on a few topics as can happen at the end of a new cycle. Arrival rates are slower and more difficult to characterize in the scientific literature than in the news. These issues make modeling scientific research trends prohibitive. Despite the limitation of existing technology, emerging topic methods have proved somewhat useful in this area, though emerging topics are discovered at a very coarse-grained level [5]. The burst detection algorithm [6], developed by Kleinberg used widely for emergent topic detection in the news [7], has been adapted in this area [4], [5]. This method identifies bursty or high intensity states of topic frequency by tracking changes in the arrival rate of topics in a stream.

Merely identifying bursty or emerging topics does not sufficiently address the needs of the scientific community. Bursts have many characteristics, and the importance of a burst to a specific person will be based on the value of each characteristic. First, science is cyclical. Research topics can reemerge and what may seem like a new topic could be merely a revisit of an old question [4]. Second, bursts can occur at varying levels of granularity. A burst of activity could correspond to fifty, one hundred or thousands of documents. A burst of fifty or even less could be immensely important to researchers whose work is relevant to the associated topic. NIH program officers, on the other-hand, may only be interested in large-scale trends at the discipline

level. Third, merely identifying all bursts, without any context is unlikely to be useful to researchers. At any time there are numerous bursts for various research topics, and a given person will need a way of ascertaining which bursts are most relevant to them. Fourth, it is important to determine which bursts are likely to last, and identify them as early as possible. Bursts can be at a high frequency level for long periods, with steady increase in intensity. Conversely, some topics are only relevant for a short period of time. What is really needed is a way to determine which emerging topics are likely to stay bursty. Also, if a topic has been bursty for a long time, one may need to ask is this trend as interesting as one that is newly in the literature, or re-emerging?

Though emerging topic detection could be performed for any research domain, I focus on biomedicine for several reasons. First, in biomedicine, curated article information and numerous classifications systems are made freely available by the National Library of Medicine, (unlike many other research areas). Conducting a comprehensive study with sufficient data is, hence, more easily achievable in this domain. Second, in biomedicine, interdisciplinary research is conducted in a more deliberate manner than almost any other field. Translational medicine, a methodology that bridges gaps between basic and clinical science, has gained attention by both the medical research community as well as government funding agencies. Both groups identify a need for coordinated efforts to assist researchers in translating basic science research to clinical for the end goal of improving patient care. The signal-to-noise ratio problem is a significant impediment as identifying what of basic science can or should be translated is a difficult task for clinicians [8]. In addition, the U.S. approach to clinical and translational research (CTR) has been characterized as fragmented, slow, and expensive [9], [10]. This makes identifying emerging topics at the discipline level significantly important – allowing researchers in a

particular field to keep up-to-date to translational trends in their own field, and be better able to identify trends in other related fields. Third, this domain was chosen because patient care can be significantly affected by the lack of knowledge of what is current. A framework that is as comprehensive and detailed as possible is a must for this domain, and it is likely it would be more widely used in this domain than any other.

The framework, described in this work, will benefit policy makers. An emerging topic detection application could assist with both the selection of research profiles to fund, and for evaluation. An emerging topic detection algorithm could help in many ways. For example, decisions about whether to fund research based on the degree of stability of associated research topics for a high-risk project can be more efficiently made. This could potentially reduce the number of projects that are too high-risk. Early signs of burstiness could be used to determine the extent to which the research is gaining traction. That could boost a project's perceived viability. Emerging topic detection would be most useful in the evaluation stage. As noted above, there does not currently exist a method for quantifying transformative research. Performing burst detection analysis on a hierarchical topic structure has the potential to provide that. By modeling the emergence of a topic and its growth across the hierarchy, a given research area's transformation of the landscape can potentially be assessed. This will be discussed further in chapter 6. An emerging topic detection application could help NIH reach its goals. For instance, emerging trend detection could help evaluate interdisciplinary research proposals to identify novel research that has gained traction. It could help with research assessment by determining whether emerging trends in basic science are getting projected into clinical science. Large-scale interdisciplinary big science research proposal include the coordination of complex research areas, and a tool to determine novelty would be useful in assessing their potential.

The primary goal of this work is to provide a framework for identifying and understanding the significance of bursts within scientific literature by accounting for the development of topics, and the fitness of those topics. This research is meant not just to identify bursts but also to help different interest groups find important bursts early. Important bursts are those that are likely to stay bursty, be associated with somewhat novel or reemerging topics, and be at specific levels of granularity. This framework could then be used in the future to develop a burst detection application that could be used for scientists, administrations and governmental groups.

The main components of this burst detection framework are:

- A hierarchical topic structure so that topic relatedness can be used to improve burst identification. This will make it easier to develop a topic granularity model.
- Specific characterization of bursts, and generalizations of burst structure at various levels of granularity.
- Archival measures combined with fitness measures used in conjunctions with Kleinberg's method.
- Noise Reduction of result-set. For each weighting scheme, the same method is used to reduce noise. Level of granularity is used, as well as burst strength, and the impact of modifying results with h-index and impact factor. The top bursts for each measure are selected.
- Evaluation of the effectiveness of this framework based on how well modified methods identify bursts early as compared to Kleinberg method and how well my method

identifies topics which truly had a significant change in frequency for the given period in real-time.

CHAPTER 2 RELATED WORK

2.1 Research Policy Making

For decades, the progress of science has been propelled by funding decisions of policy makers like the National Institutes of Health (NIH) and the National Science Foundation (NSF) [11]. In the scientific literature, bibliometric methods have been used to both evaluate research initiatives, and for effective resource allocation by those institutions [12]. However, peer-review remains the most common method used, for those tasks, by funding agencies. Limitations of the peer-review methods, and the cost of process has made many suggest that more comprehensive biblometric methods need to be developed [13]–[17].

NIH and NSF have tried to find innovative, cutting-edge research to fund for decades [18], [19]. Wagner et al. conducted a retrospective study on NSF small grants for exploratory research [16]. The aim of the 16-year NSF program was to identify and fund transformative research. Transformative research alters the paradigms of a research area through innovative techniques. It is high-risk and high-stakes. It is high risk because it proposes ideas sufficiently divergent from accepted evidence as to seem transgressive, and is potentially impractical. It is high-stakes because it has the potential for high impact both in academia and society. It was deemed necessary to create an initiative to support this research type, because of the belief that the peer-review journal process is biased against it. Several aspects of this initiative are worthy of note. First, identifying transformative research is difficult. It can only be truly identified in retrospect. Therefore, one characteristic of programs to support it is that they will inevitably fund research that has little to no impact or long term utility. Second, there is no clear set of requirements upon which quantifiable assessment measures for transformative research should be based. These two

caveats mean that at the resource allocation stage and the evaluation stage, policy makers could benefit from tools to help them make decisions.

NSF is not alone in this focus on innovation. Beginning in the early 2000s, the National Institutes of Health (NIH) began a process now referred to as the "NIH Roadmap" [20]. The main purpose was to define a set of priorities to address current pressing scientific challenges. Three major themes emerged from this endeavor; among them was *New Pathways to Discovery.* This theme is focused on the use of genetics, molecular and cell biology to create innovative toolkits for biomedical researchers. Another was *Research Teams of the Future* – developed based on the need to solve present-day complex social, medical, technical and environmental problems using knowledge integrated from a diverse set of disciplines. This focus on research teams or mega-science is focused on using cutting edge-technologies for the purposed of knowledge production [11], [21]–[23]. These endeavors are very important as output of novel medical research has declined in recent years due to the difficulty in translating basic science to clinical research [24].

Using bibliometrics to assist policy makers is not a novel exercise. Rafols et al. generated overlay maps to assist policy makers visually locate bodies of research that do not fit into traditional disciplinary boundaries [15]. Ordonez-Matamoros et al. provided an example of national level team performance assessment with bibliographic methods as a response to policy to encourage team-based research [25]. Ponomariov et al. [26] used bibliometric methods to analyze the effects of research institutions as the institutional policy response to technical and scientific demands. These large-scale analyses demonstrate the ability of bibliometric methods to identify large-scale structural trends that could not feasibly be performed manually. Bibliometrics are not only effective in those cases, however. Compared to bibliometric methods,

there are many limitations of the peer review process used for policy decision-making[12]. Peer-review suffers from subjectivity, and high-cost. The time that experts take to evaluate portfolios is time away from other research. Bibliometric methods have proven efficient at a guaranteed lower cost. Though bibliometric methods have advantages over peer-review, peer review is the more common method used for research evaluation, and resource allocation by policy makers. There are potentially many reasons for this. For instance, using bibliometric methods requires a specialized skill-set, and the necessary applications linked to data may not be available.

On the other hand, there are also issues inherent in bibliometric methods that make applying them problematic. For instance, ranking-based methods, which are often used for these tasks, lack the interpretive flexibility that an overlap map has [15]. Bibliometric methods are not silver-bullet policy decision-making tools. They will be most useful when the information they present is rich and dynamic enough to support decision-making.

2.2 Use Cases for Researchers and Clinicians: Medical Literature Search Tools

In addition to policy makers, clinicians and researchers could benefit from an emerging topic detection application. Researchers could identify updates in both their main area of research and, in related areas. Emerging topic detection would be extremely useful for researchers aiming to write reviews. For clinicians, tools that help them keep up to date in their discipline would be extremely useful.

Biomedical researchers are tasked with solving intricate problems. Complex problem solving requires planning and an understanding of the relationship between many concepts, which is very difficult without external aides. Studies of the information seeking behavior of scientists suggest

that researchers are often confronted with what is referred to as *weak problem solving*, associated with a vague understanding of the problem space, an inability to come up with a systematic plan to resolve information needs and a lack of prior domain knowledge [27]. These issues can be exacerbated by the difficulty of determining what is considered the most current in a given research area. A researcher's career prospects are contingent on how impactful his/her research is to the development of scientific knowledge [28]. If researchers can assess whether a given research area's topic occurrence rate is declining or increasing that knowledge can help them make decisions that will affect the impact they have on their field

Clinicians need skills and tools to find evidence at point of care. Numerous tools exist to summarize, or improve search of Medline citations for point of care discovery [29]–[35]. Clinicians have very limited time to answers questions during a typical day [36], [37]. Most clinicians rely on summaries and practice guidelines regardless of whether these resources are evidence based [38], [39]. Those who study these behaviors are calling for information tools, which alert clinicians to new, relevant or valid information[40]; tools which tailor information to the appropriate specialty of each physician. Such tools could be greatly improved with the use of a robust emerging topic detection framework, as it would allow for the identification of new ideas gaining traction in the research community.

2.3 Topic Detection and Tracking

Related to the area of emergent topic detection is topic detection and tracking (TDT). In the topic detection literature a topic is defined as a set of items strongly related. In the context of news, a topic is a set of news items strongly related by a special event. Topics can also be defined in terms of about-ness. Cluster analysis in topic detection is the process of grouping

related items into bins that represent a topic. For TDT, clusters are created and updated based on documents arriving in a document stream. Topic detection and tracking research is concerned with two major tasks [41] :

- Segmenting a stream of data into distinct stories.
- Given a sample number of sample news stories about an event find all the following stories in a stream.

One approach is that used by Eichmann et al.[42]. First, the TF-IDF weighting scheme is used to perform cosine similarity to find related documents:

$$\mathrm{tf} - \mathrm{idf} = \mathrm{tf} * \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

where TF is the term frequency in a document, D is the number of documents in a specific collection, and $\{d \in D : t \in d\}$ is the number of documents in which term t appears. IDF ($\frac{|D|}{|\{d \in D : t \in d\}|}$) is the inverse-document- frequency of a term within a larger document. TF-IDF boosts a word's importance to a document based on how important it is to all documents. So, if a term appears a few times in a document, but is rare in the corpus it will have a higher TF-DF score than if it appeared in many documents. After TF-IDF is computed for each term for each document, cosine similarity is calculated pair-wise between documents:

$$\cos(d_i, d_j) = \frac{I * J}{||A|| * ||B||} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (A_i)^2}}$$

where A and B represent the TF-IDF term vectors for document d_i and d_j respectively. Cosine is calculated by multiplying the dot product between the two vectors, and dividing that by the

multiplication of the magnitude of each vector. Once cosine is calculated, documents can be grouped together into clusters using the pair-wise similarity scores between documents.

Since one is dealing with a stream of data, one can choose a point at which term weights are updated [42]. At each iteration, a topic must be within a certain threshold of similarity to a given cluster to be clustered at all. Items not clustered in one iteration may be part of a cluster in a future iteration. This method makes it possible to account for the need to update the categorizations as new items come in. However, if items are added to existing clusters as they arrive in the document stream, bias can be introduced because new items cannot influence the *formation* of clusters as the older documents did. The order of cluster merging is then different from what it would be if all documents were considered in the first iteration.

Another method used for topic detection is the Dragon method [41]. The distance measure used is Kullback-Leibler:

$$\sum_{n}\left(\frac{s_n}{S}\right)\log\frac{\frac{c_n}{S}}{\frac{(c_n+s_n)}{C+S}}+\sum_{n}\left(\frac{c_n}{C}\right)\log\frac{\frac{c_n}{S}}{\frac{(c_n+s_n)}{C+S}}$$

where c_n and s_n are story (i.e. topic) and cluster counts for word w_n. For each subsequent iteration, the algorithm considers switching each item to another story based on the same measure as before. This algorithm is then able to account for bias that is introduced using comparisons between new documents and clusters of early documents.

To evaluate TDT algorithms, two errors are considered: misses (in which the target event was not detected), and false alarms (an event was detected but it is not valid). Precision and recall scores are typically used for evaluation [43]. Precision is the proportion of retrieved items for which the target event was detected, and recall is the proportion of target events that were

detected. TDT tasks are usually tested on datasets for which class labels are known, so, this is not an unsupervised learning problem, and typical clustering evaluation methods do not need to be used.

There are several aspects of this problem area that are relevant to emerging topic detection. First, before one can determine whether a topic is bursting, topics must be identified and documents associated with them. Second, topics exist at different levels of granularity. Topics can be broad and cover a wide spectrum of data. Within broad clusters, there could be many nested clusters that can only be identified by sub-cluster identification, or by changing the threshold of similarity to determine cluster membership [44]. There are many difficulties associated with TDT. Among those are determining ideal cluster parameters to identify meaningful clusters, and to characterize those clusters appropriately [45].

2.4 Kleinberg's Burst Detection

One method for identifying emergent topics in a stream of data is Kleinberg's burst detection algorithm. This algorithm, principally used to analyze news timelines, detects sharp increases in occurrences of topics over time by analyzing their arrival rates. It has become a popular method in analyzing both news and social media trends [46]–[50]. This method has also been adapted for use in detecting scientific research trends [4], [51]–[53].

Kleinberg's burst detection algorithm [6] can be used to find moments of intensity for a topic amidst the noise of unrelated documents in a text stream. This model uses a probabilistic automaton to model the conditional prior probability of each potential state of intensity (q) based on the arrival rate for each item (x) associated with a given topic:

$$\Pr[q|x] = \frac{Pr[q]f_q(x)}{\sum_{q'} \Pr[q']f_{q'}(x)}$$

where t represents a particular time interval, b is a scaling parameter; PR[q] is the probability of q. Gap x between messages i and i+1 is emitted in a probabilistic manner distributed according to the exponential density function $f(x) = \alpha e^{-\alpha x}$, for a parameter $\alpha>0$. This means that the probability that a gap exceeds x is equal to $e^{-\alpha x}$. The parameter α can be referred to as the rate of message arrivals. To better understand how this method works, consider the following scenario. Imagine an automaton A with two states q_0 and q_1, which correspond to a low, and high state respectively. When A is in the state q_0, messages are emitted at a slow rate. When A is in state q_1, messages are emitted at a faster rate. Between messages, A changes state with a probability p, and with the probability of staying in its current state with the probability 1-p.

There is a cost for jumping to a higher intensity state from a lower intensity state. Finding a state sequence that maximizes the previous equation is equivalent to minimizing the following cost function:

$$c(q|x) = b \log\left(\frac{1-p}{p}\right) + \left(\sum_{t=1}^{n} - \log f_{i_t}(x_t)\right)$$

The first part of this cost function favors sequences with few transitions, and the second part conforms well to the sequences of gaps. Using this model, one can define a burst of intensity j to be the maximal interval over which state sequence q is in a state of index j or higher. More precisely, it is the interval [t, t'] so that $i_t, \ldots i_{t'} >= j$, but i_{i-1} and $i_{t'+1}$ are less than j. Using this model, one can also identify the natural nested structure of bursts. Sub-intervals that are bursts

of intensity j+1 may be contained in the burst of intensity j. This method can be used to find terms that exhibit the most prominent rising and falling patter over a limited period of time.

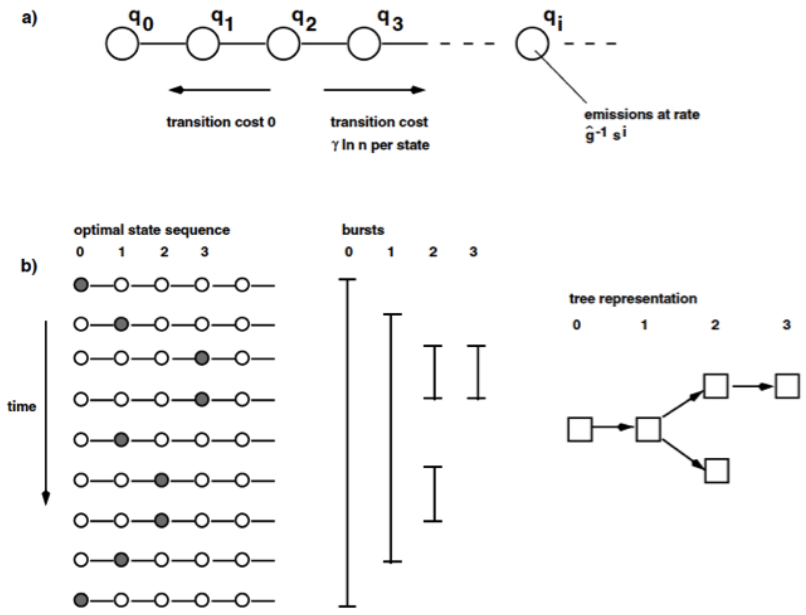


Figure 1 Kleinberg's hierarchical representation of bursts [6].

Figure 1 presents an example of the hierarchy of bursts. The use of this model leads to identifying bursts with clear beginnings and endings. By focusing of the boundaries of burst intervals one can determine the extent to which state transitions are "sharp."

2.5 Probabilistic Burst Detection Different from Kleinberg's

Morinaga et al. developed a finite mixture model to detect bursts [54]. Mixture models are used to make statistical inferences about the characteristics of sub-groups within a larger set. A mixture model corresponds to the mixture distribution that represents the probability distribution of observations for the entire set. For the purposes of burst detection analysis the full set is the set of documents, and the sub-groups to be detected are the topics to be found. K is defined to be a positive integer representing the different topics. Morinanga et al. assume each text has only one

topic and a text having the ith topic is distributed across the vocabulary space according the probability distribution with a density: $p_i(x|\theta)(i=1,2,\ldots K)$, where θ_i is the real-valued parameter vector for each topic. Each document d is distributed according to a finite mixture distribution with K components:

$$p(x|\theta{:}\,K) = \sum_{i=1}^{K} \pi_i p(x|\theta_i)$$

where $\pi_i > 0$ and $\sum_{i=1}^{K} \pi_i = 1$. π_i denotes the likelihood that the i-th topic is to appear in a given text stream. The finite mixture model allows for the representation of the topic structure of a set of documents, something not achievable with the Kleinberg method. In terms of their model, topic structure is identified by A) the number of components K, B) the weight vector $(\pi_1,\ldots, \pi_k)$ and C) the parameter values. Topics must first be characterized by classifying each text into the component for which the posterior is largest, and then extracting features terms which best characterize the classified text.

Components are selected by starting out with a large set of components, and dynamically reducing those based on how well their results distribute documents across the components based on the topic structure described above. The method was tested on a help-desk dataset, and the authors claim that the algorithm is able to detect emerging topics in a timely manner. However, they do not measure against any standard, or compare their results from those from other methods.

The work of Chen et al. uses Expectation Maximization for emerging trend detection. The Expectation Maximization (EM) algorithm is an iterative optimization method consisting of two steps [55], [56]. This method is used to estimate the maximum likelihood of an event, when there are unknown probabilities that must be accounted for. It is used when distributions are not well

behaved, or there are too many parameters. The method uses two variables: observed variable x, and hidden variable y which generates x. The method assumes two probabilities: $P_\theta(x)$, and $P_\theta(y)$. θ represents the set of all parameters of the distribution. The two steps of the algorithm are:

1. Expectation step: Compute expectation of $\log(P_\theta(y,x))$. This step tests different sets of distributions of hidden variable y and finds the probability that it generates observed variable x.
2. Maximization step: Find y distribution that maximizes the likelihood of x.

In this space, EM algorithms are useful for several reasons. For clustering based methods, EM is effective because they make it easy to compare clustering results based on log-likelihood levels, and it can be used to make predictions of cluster membership.

Chen et al. tested their method on three datasets retrieved from *Web-of-Science*: Social Network Analysis (1992-2004), Mass extinction (1981-2004), and Terrorism (1989-2004). Each dataset included article title, abstract, and citations to existing articles. They used CiteSpace, a program implemented in java, to analyze and visualize citation networks. Using this application, the authors were able to select a time interval for analysis, and then partition that interval into equal-length intervals referred to as time-slices. The networks can then be merged. The following attributes were compiled for each article and used in the EM-step: citation counts throughout the entire time interval, betweenness centrality, the first author of the article, the year of publication, the source of publication, and the half-life of the article. The half-life of an article is the time at which an article has received 50% of the citations it will receive in its lifetime. Betweenness centrality is a connectivity measure based on how many shortest paths from all nodes to all others that pass through a given node. It gives an indication of how central a node is to the network.

Network nodes are then clustered by these attributes using the EM algorithm implemented in the Weka open-source data mining toolkit.

To evaluate their method, they tested the ability of their algorithm to identify emerging terms in a datasets where the paradigm-shifting terms to discover are already known. Their methodology was successful in identify those emerging terms. However, how it would perform in an open domain setting where the emerging terms to discover are unknown was not determined. After running their analysis they found that betweenness centrality was the only attribute that really accounted for global structure and had a significant impact on how well their algorithm performed. That the other measures could not be made useful is important to note. They may account for the importance of a node, but not its place in the network.

2.6 Physics Inspired Burst Detection

He et al. use concepts from physics to characterize burst of Medical Subject Heading (MeSH) terms [57]. The authors expand on the idea of burst detection by incorporating physics concepts such as velocity, acceleration and mass. The authors provide several criticisms of Kleinberg's method for scientific burst detection. First, they claim the underlying arrival process may not be clearly Poisson, an assumption made by Kleinberg. Gaps between occurrences could be wide for instances. Second, the number of publications in PubMed has increased in recent years, which means that it could appear that the arrival rate is higher even if the overall proportion of documents on a particular topic remains the same. Third, the Kleinberg model is memory-less; future state depends on the present state not past state. However, He et al claim that in the case of scientific topic emergence this is problematic. The probability of a term reoccurring in the literature is not independent of all the past times it occurred. The more popular a topic already is

the more popular it is more likely to be. This is referred to as the Mathew effect, and it has been demonstrated to be true when one considers the popularity of a document (where popularity is defined by the number of citations), and the prominence of an author (where prominence is characterized in terms of productivity and impact) [58], [59]. Fourth, and this will be discussed further below, science articles are published at established intervals, and do not arrive at sporadic intervals. Therefore, the authors state, an arrival-rate based algorithm is not the most appropriate.

Instead of characterizing bursts in terms of the arrival rate they characterize them in terms of changes in momentum. Using the concept of momentum makes sense, because each research area can have its own maximum level of intensity. Also, some topics just stay bursty. Consider cancer – at any given time there is a constant increase in documents concerning this topic. The Kleinberg method may flag it as bursty when it is really a commonly studied topic. He et al. borrow terms from classical mechanics, such as mass, position, and momentum. In physics momentum is the product of mass and velocity. Heavy objects have greater momentum than lighter ones going at the same speed. The greater the mass, the harder it is to bring an object to a stop. Position in this context refers to the count of documents at any given interval. Mass refers to the current importance of the topic, which can be inferred from article citation counts, journal impact factors, and journal relevance measures. The idea behind using momentum as a metaphor for importance is that prominent topics should be more likely to stay up to a particular "speed" for a longer period of time. A burst is a period where acceleration of velocity is positive. Identifying acceleration is achieved by taking the second derivative of positions across a time-line. According to their definition, a topic "burst" is an interval of positive acceleration.

Adapting the data to popular stock market trend analysis techniques, such as Exponential Moving Average (EMA), circumvents the difficulty of measuring momentum directly [60].

Moving Average (MA) is used to smooth out the noise of price in stock market analysis. He et al. [57] adapt it to the problem of scientific emergent topic detection to smooth out the noise of gaps in the arrival rate of topics. It is a moving window in which to calculate the average. In financial settings, a simple moving average (SMA) is the un-weighted mean of the previous *n* data points. An exponential moving average (EMA), also known as an exponentially weighted moving average (EWMA), applies weighting factors, which decrease exponentially. The weighting for each older data point decreases exponentially, though it never reaches zero. He et al. use moving average convergence/divergence (MACD) to estimate velocity, which is a type of EMA. This method takes the difference between moving averages for consecutive intervals.

The authors generate MACD histogram, which is the second derivative estimate, to indicate positive acceleration. It takes the third day moving average, and the MACD from the previous interval. They define bursts as a kind of linear filter. Given a two-MACD histogram the difference in their parameters corresponds to the differences of specific burst intervals. This allows for a better definition of bursts, according to the authors, as the Kleinberg model identifies periods of intensity, and not necessarily increases in the rate of occurrence of a term.

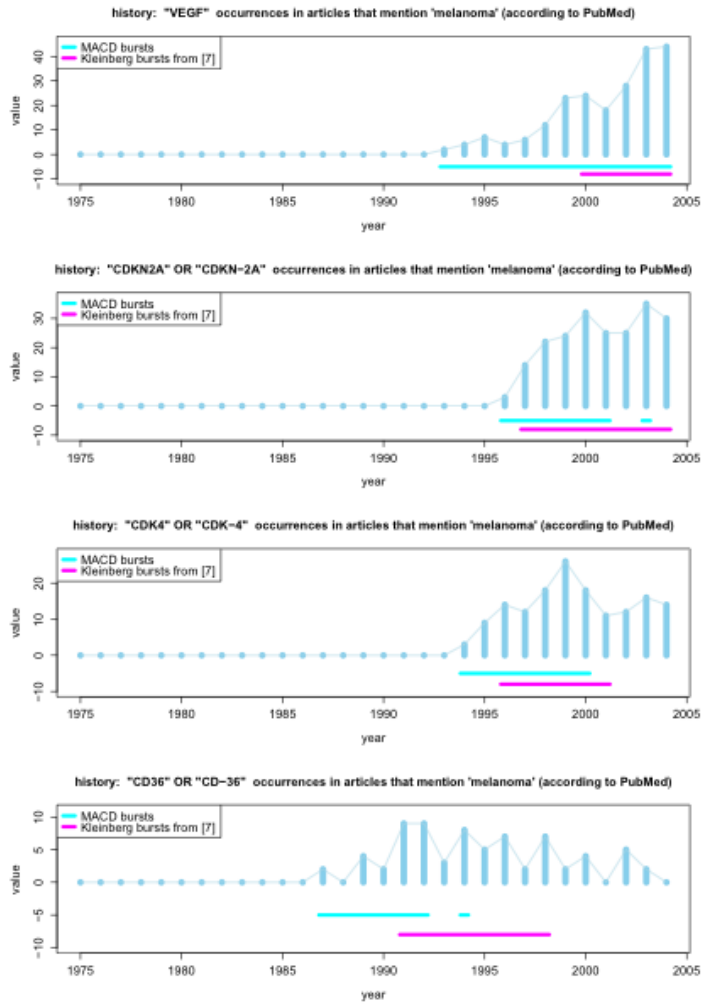


Figure 2 Result of He et al.'s comparison with the Kleinberg method for a select set of gene names [57].

To test the accuracy of their model, they test for the detection of major developments in science, such as the flurry of genetic research in the 90's after the human genome product. Their model does not actually identify momentum. They do not incorporate their notion of mass into their algorithm. They did not choose an example to test their model that demonstrates how well it performs on topics for which trend characterization is unknown. Rather, they chose what appears to be an easy example, as they are running their model on what is among the most bursty topics in scientific history, during a period when there are constant publications on the given topic. Their results are displayed in figure 2.

When compared with the Kleinberg model, their method did a better job detecting bursts earlier in the time-line. Therefore, it could be successful at predicting oncoming bursts, or predicting a high level of intensity before it happens. However, their methodology could be improved by viewing the maximum velocity of neighboring topics. Positive acceleration by itself will not identify the most important bursts. A slight increase in momentum is not always going to represent a truly trending topic. One of the advantages of the Kleinberg model is that it makes use of a threshold of intensity for each state. That makes it easier to compare bursts and assess the popularity of a given trend. Figure 2 shows that the burst period for the Kleinberg method ends later. A topic can experience a decline in popularity, while still being relatively popular.

He et al. claim that a model that analyzes the arrival-rate of a topic is inappropriate for this task because the interval of scientific publications is standard, and discretized already. Scientific articles are published at specific intervals. However, there are two problems with their assumptions. First, not all journals publish at the same interval. Some have publications every month, some every three months, and others even fewer times a year. Between the intervals for one journal, many other articles are being published. Second, when the burst detection algorithm is operationalized, an interval of time must be chosen. Within that interval there could be many instances of a topic, which arrive at slightly different times. However, because of the difficulty in representing exact times, the Kleinberg algorithm must treat them as all arriving at the same time. So, the way the arrival rate is represented is the same whether one is talking about news items, for which arrivals are more sporadic, or scientific documents for which arrivals are regular. Choosing an interval in which there are likely to be occurrences makes detecting the arrival rate easier. Consider an example from news. There may well be periods during the day in which there are very few news stories. If one is interested in detecting bursts, one will set up the interval to ensure

that there is the opportunity for occurrences of the most important news items within each news interval. For instance, say that between 11:00am and 12:00pm there are typically no news stories and again at 3:00pm to 4:0pm there are few news stories. If an interval of an hour is chosen to group items appearing together, there will be fluctuations in the trend-line for most topics. On the other hand, if a larger interval were to be selected the trend lines would fluctuate less as each interval would correspond to greater activity. This will make it possible to detect the highest arrival rates. Choosing an interval that will smooth over gaps of low activity makes it possible to have arrival rates with the smallest between arrival times. Also, He et al. use concepts from physics that are inappropriate given their claims. These concepts are meant for continuous, not discrete data, and therefore do not solve the interval problem. The estimates are inappropriate for discrete datasets when the gaps are large. Further, momentum can itself be characterized as the arrival rate of a topic within their framework. Therefore, velocity is very similar in concept to changes in arrival rate.

Their model appears to be more effective than Kleinberg's detection model because they are able to identify bursts earlier. This is deceptive. The burst detection model identifies bursts that are of certain intensity. He et al.'s methodology is focused on changes in momentum. A change in momentum can be very slight. Though they are able to identify bursts early on, their method would identify momentum of topics which were merely experiencing a temporary fluctuation. There is no way to know how much more noise would exist in their results if they used an open domain problem.

He et al. claim that an advantage of using moving average is that it smooths over gaps. However, if the interval of time is selected correctly, gaps should not need to be smoothed over as they reflect the actual fluctuation in interest many topics experience. Consider the following. Say

I have two data points in interval 1, none in intervals 2 and 3, and two data points higher than interval 1 in interval four. Their model would make it appear that there is a constant increase because it would smooth over the gap of the two intervals for which there are no values. The Kleinberg method will actually show the fluctuation and the topic will appear to have a more intense burst in the fourth interval than would be identified in He et al.'s. It is in fact possible that the latter's model will not even identify a change in acceleration, because the smoothing over may end up making momentum appear steady.

One other major limitation of He et al.'s evaluation is the use of the MeSH hierarchy. According to the developers of this classification system, treating it as an ontology is inappropriate[1]. It provides descriptive terms (words and/or phrases) that are useful when searching for categories of documents. The hierarchy is used to provide context. In other words it is not a conceptual hierarchy meant to give an overview of the disciplines in medicine. It does not provide breadth and scope of fields in terms of topical relatedness. For instance, if term b appears under a term it does not mean that b is part of a subset of a. It means that b is considered in context of a. B could be a drug that is used to treat disease a, and the terms are related but not in the way they would in a true ontology. Therefore how terms are grouped using MeSH can be problematic.

Another limitation of using MeSH for burst detection is it can't detect new terms as they emerge. This is an important problem in this space. New terms are added to this classification system retrospectively, not when they appear, and therefore any burst detection system using this methodology will not be able to account for terms not yet in MeSH. Lastly documents are only

[1] I called the MeSH office July 12 of 2011, and discussed this issue with the developers.

tagged with MeSH terms once. If a new term appears that is relevant to an old document, that document's metadata is not updated.

Moerchen et al. use a similar method based on the change in the relative frequency terms in biomedicine [61]. They evaluate their method by testing how well they are able to predict the inclusion of new MeSH terms. There method suffers from the same limitations of He et al.'s.

2.7 Citation Based Burst Detection

Because of the difficulty in identifying and tracking emergent areas through raw text, the study of this phenomenon has typically been accomplished by tracking shifts and progress in a known new area using citation networks.

Novel research produces changes in the structure of citation networks. Linkages between unconnected or loosely connected areas are generated as emergent research can help re-cast, reformulate, and extend previous research. New findings also create their own clusters of documents. The detection of evolving topics in a particular research area, based on shifts in citation patterns over time, has been studied extensively in the literature [62]–[67]. The goal is to track the research front, i.e. the topic area and people involved. This is typically achieved by using citation-based methods, such as co-citation analysis and bibliographic coupling. Mapping science can assist policy makers and funding agencies keep track of disciplines, and make informed decisions when determining which research areas to support.

Analyzing the research network, and attempting to identify topic drift may improve the ability of algorithms to identify important bursts. Small's model of scientific development [68] is

useful for understanding the relationship between burst detection as detected from citation networks and identifying seminal papers:

- Upon the nascence of a research area, discovery papers attract many citations.
- A series of papers extending the ideas of the original papers appear--becoming a group of highly cited papers.
- A series of papers appear that are cited at a more "normal" rate. Citation rates for the discovery papers decreases.
- Normal production of papers continues until the research area becomes obsolete, whether or not it is replaced by a new area.

Analysis of the distributions of papers at different stages of the research front bear this out [65]. With this model in mind, bibliographic techniques can be used to analyze the citation network to identify bursts and seminal papers.

Morris et al. use bibliographic coupling to group related documents [62]. Data from ISI was used to map several research fronts based on identifying coarse-grained keywords. Bibliographic coupling uses the reference lists between documents to determine similarity:

$$S_{ij} = \frac{bc_{ij}}{\sqrt{N_i N_j}}$$

where bc_{ij} represents the number of references paper i and paper j have in common. N_i represents the number of references for paper i and N_j represents the number of references for paper j. Once

similarity between documents is determined a clustering algorithm is used to group documents together. In this area agglomerative hierarchical clustering is often used [69].

Agglomerative hierarchical clustering is a bottom up approach. In the initial phase each document is a cluster. Similarity scores are calculated for each pair-wise cluster. At each step, the clusters with the highest average score are merged, until a specific number of clusters is identified or the threshold of similarity is met. In Morris et al.'s work this method was extended with Ward's linkage algorithm, in which merges are done such that variance in inter-cluster distances is minimized.

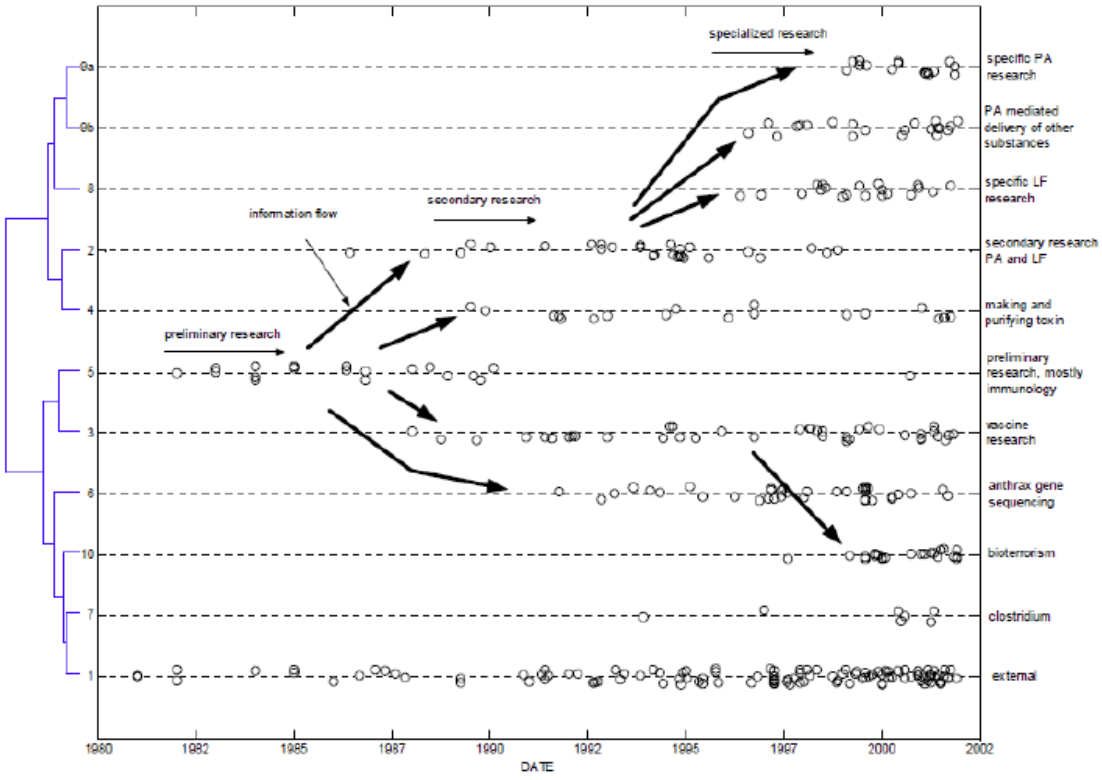


Figure 3 Morris et al.'s document timeline for the anthranx dataset. Information flow, derived from exploration of citation patterns, is shown as heavy arrows [62].

Document titles were manually inspected to find appropriate cluster labels. Some clusters

were subdivided based on manual inspection. Information flow between clusters across time was identified by exploring citation patterns. From figure 3, from Morris et al.'s work, one could identify seminal papers by analyzing the flow of information across time.

Takeda et al. use a slightly different method and analyze the clusters to identify emerging topics [63]. Their method makes use of inter-linkage and clustering based on modularity

$$Q = \sum_{s=1}^{N_m} \left[\frac{l_s}{l} - \left(\frac{d_s}{2l} \right)^2 \right]$$

where n_m is the number of clusters, l_s is the number of links between nodes in cluster s. d_s is the sum of degrees in cluster s. Modularity measures the relationship between the connectivity of nodes within groups with the connectivity of nodes between groups. A modularity based clustering method optimizes for cluster density.

Cluster labels were generated with a bottom up method. Topics were first identified from sub-topics. Then these topics are aggregated. Once the clusters were identified, the authors generated a timeline of the publications for each cluster. Not all tightly coupled documents constitute a research front. Therefore, they identified average publications by year, and manually identify emerging trends. They found that as the research network emerges, it is characterized by loosely related clusters.

If one were to combine the approach of Morris et al.'s with Takeda et al.'s one can see there is potential for both identifying topics bursts, and finding seminal papers from those bursts. The semantics of similarity is different for each method. For Morris et al., clustered documents cite the same papers. In Takeda's method, nodes are linked because one cites the other.

Identifying seminal papers using the former method requires identifying information flow between clusters, whereas for Takeda's method one would need to identify information flow between nodes in a particular cluster as well. A true research front would also need to be identified. Assuming one identified emerging research papers how effective would these mappings of clusters across time be in identifying seminal papers.

de Salla demonstrated that scientists tend to cite recent papers [70]. Therefore, the most relevant seminal papers may not be heavily cited. Consider Small's model of scientific development mentioned earlier in this section. He states that many papers do not cite the earliest papers, but typically more recent ones. One can infer that many of the papers that might seem seminal are actually not the discovery papers, but are the second phase of highly cited papers. de Salla found that papers published in 1961 cite earlier papers at a rate that falls off by a factor of 2 for every 13.5-year interval measured backward from 1961. Thirty percent of papers cited are between 1 and 6 years old. Will the seminal paper actually be the one cited? This has been a question that has been asked by prominent researchers in the area of bibliometrics [71]. During the period between 1961 and 1975 Watson and Crick's famous paper presenting the double helix structure of DNA had relatively low citation counts [72]. When an area is new, it is not always well defined. Takeda et al. found that in the early stages of development research areas are defined by loosely related clusters [63]. Not only will the area lack tight connection between clusters, it will lack a standard of terminology. The may make using citation analysis for emerging trend detection more difficult.

Ahlgren et al. compared text-based methods with citation-based methods for research front detection [73]. They had experts categorize documents into groups and then tested the ability of various clustering methods to match those categorizations. The citation-based methods

performed the worst, perhaps for the reasons stated above.

2.8 Burst Detection Based on Topic Drift

The concept of topic drift has also been used to develop emerging topic algorithms. Topic drift can be defined as small successive modifications in a topic [74]. When a topic drifts, the original formulation of the topic and the changed topic share a significant amount of similarity, but the actual topic has changed. Citations-based methods to identify topic evolution have been studied in the literature [75]. There are several related reasons why this is useful for burst detection analysis. When a topic is introduced into a particular field, it flourishes only under certain conditions. There must be interest in the field, and researchers capable of dealing with it. The support it receives (in terms of related research that can support it) is dependent on time, recent work, popularity, visibility of associated journals, and conceptual models of active researchers.

One method, used by Qian et al., clusters documents and then tracks the change in cluster membership for the top nodes [76]. They first construct citation graphs using a modularity-based method. The authors defined topic emergence as the separation of the top cited paper from the old cluster and formation of a new cluster. They base their research on the hypothesis that "a new research topic must include recently most cited papers (top papers), otherwise it could not draw enough attention". They distinguish two clusters using the Jaccard coefficient to measure the similarity between cluster C_i and C_j as defined by:

$$sim(C_i, C_j) = \frac{|C_i \cap C_j|}{C_i \cup C_j}$$

The authors evaluate their method with a high-energy physics dataset from the KDD Cup 2003. They use a citation turnover value of similarity (i.e. the number of documents that eventually join to new community) instead of a using the Jaccard method for evaluation purposes. Though they don't have a gold standard to use for comparison, their results indicate that when they detect a new cluster there is an eventual high turnover rate between 44 and 88 percent to newly identified areas from the research areas they are presumed to come from.

Their work helps support the idea that one can identify emergent trends using topic drift. However, there is no actual way to determine how many bursts their methodology did not detect. What if the new topic does not cited the most highly cited node in the old community, but a medium cited one? Some bursts may be associated with topics that have hitherto been somewhat obscure and then became more prominent. More analysis would need to be conducted to test the viability of their method. Also any citation-based method is going to suffer from the problems identified with citation-based methods described at the end of section 2.4.

2.9 Co-word and/or Cluster Based Burst Detection

Co-word frequency, often coupled with Kleinberg's method can be used to identify the relationship between bursts. Mane et al. [5] use the complete set of papers published in PNAS from 1982-2001. They first chose the top 10% most highly cited documents in the entire span. They track the burstiness of each terms associated with those documents using Kleinberg's model, tracked the co-word occurrence of topics and topic bursts, and had experts review the results.

Once the most highly cited documents were identified, the team mapped their terms Figure 4 displays the trend-lines for the top 10 most important words. Importance is based on frequency, and the review of experts in the field.

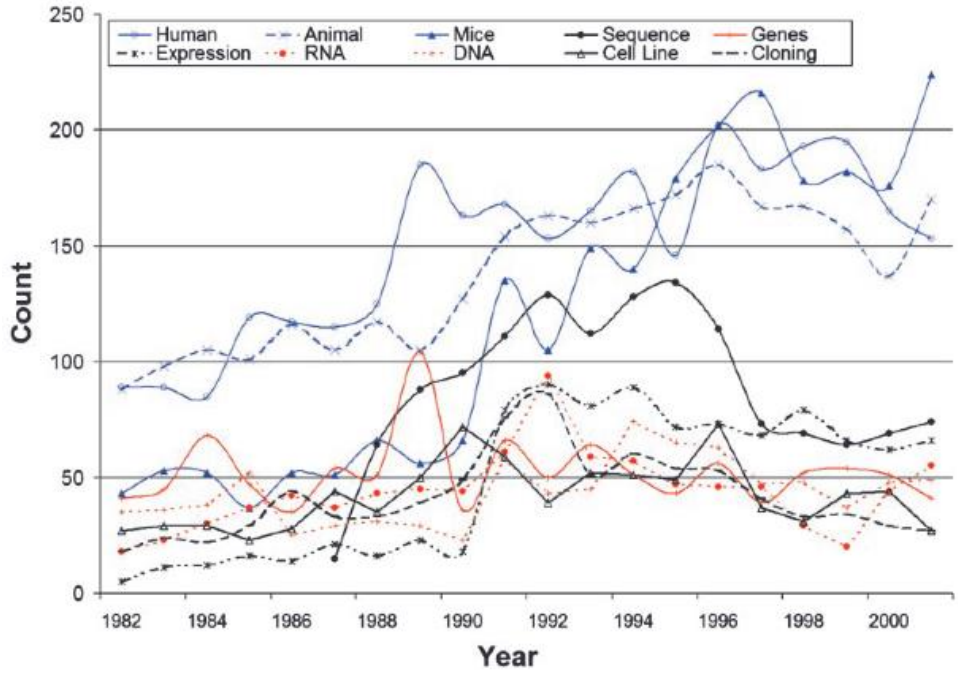


Figure 4 Mane et al.'s results for PNAS' most highly cited documents [5].

The topics discovered are all highly related and very coarse-grained. They all related to DNA. The team used raw counts, which is problematic. In the past several decades, there has been an increase in papers published in nearly every discipline in science. Relative to rate of publication for other topics there may not actually be a burst for some of these categories.

They then analyzed the topics using the Kleinberg method. For the 10% most cited documents there were 1027 words mapped and 991 experienced at least one burst, and 34 had a least two bursts. The next step in their process was co-word correlation analysis, which measures the strength of association between two words. This was done to apprehend the association between bursts. They were interested in understanding the association between the top fifty terms with a combination of high burstiness and frequency. Figure 5 represents the co-word space.

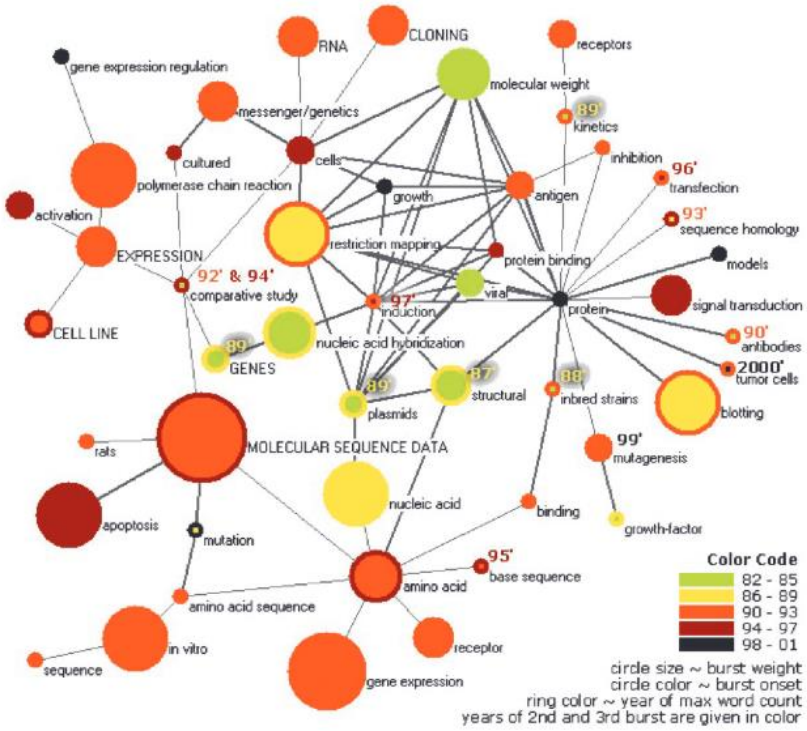


Figure 5 Mane et al.'s results of co-word analysis [5].

Figure 5 displays temporal information as well as burst strength information. The shift from the focus on structural properties of cells in the 80s to the focus of sequencing data in the nineties can be easily gleaned from this data. Their co-word analysis demonstrates that burst detection can, in fact, be used to give a very detailed overview of the progression of a field.

They found there was a low correlation with frequency and burstiness, which surprised the authors. This confusion indicates a clear lack of understanding of what a burst should look like. The most frequently used words in any field are likely to be words that are used in the most documents consistently, and therefore fluctuate the least. A burst can occur only when a word is not already common despite its frequency. By focusing on the words only at the top, the researchers have identified the trends most likely to be visible to the community and not necessarily helpful. They have provided a nice overview of the history on the PNAS journal, but have not demonstrated its utility for emergent topic detection in real time. They created a methodology to identify not what is new, but what is most dominant.

Some researchers use text-based clustering to detect emerging trends [77], [78]. Terms are clustered in each time step and then clusters are connected across time. Clusters that cannot be connected to clusters from a previous time period are candidates for new ontological concepts. . While this methodology may make it possible to find strictly new terms there is no representation of the structure of its topic growth, and it is therefore limited as a methodology for burst detection.

2.10 Burst Detection Using Eigen-Trends

Eigen trends have been used for burst detection of scientific research. Chi et al. uses eigenvectors to identify trends [56] . An eigenvector of a square matrix A is a non-zero vector *v*, that when multiplied by *v* yields a constant multiple of *v*. Many of the values will now be zero and or will change such that only certain values will appear large or small and that is believed to correspond to a particular portion of the dimensional space that can then be analyzed. It is a feature identification and reduction method. It is based on time-series derived through single-value decomposition of an arrival matrix:

$$D = \begin{matrix} d_{11} & d_{12} & \ldots & d_{1n} \\ d_{21} & d_{22} & \ldots & d_{2n} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ d_{m1} & d_{m2} & \ldots & d_{mn} \end{matrix}$$

where n is the number of intervals and d_{id} is the number of documents at interval i, and m is the source (e.g., country of origin). Single value decomposition is a method that transforms a matrix into simpler meaningful pieces. Through single value decomposition the matrix D can be recast into the multiplication of three matrices $D=USV^t$. The first eigen-trend is most important as it correlates to the importance of an individual source for a particular trend. Chi et al. (44) showed

eigen-trends can be more sensitive to the over-all contributions of high-authority sources while less affected by noise.

A trend prediction index can then be generated from the above trend matrix. One method is to identify the slope of the linear regression line that bests fits the data. This is the most commonly used prediction method in statistics [79]. Evaluating the success of the trend prediction index is a widely used appraoch in TREC evaluations [80]. This approach outputs two metrics. The first is precision rate at recall (Pre@R). This metric denotes the precision rate at the r-th position in the ordering. Precision rate refers to the relevance of the articles retrieved. Recall relates to the percentage of relevant articles that are retrieved. So Pre@r is r/R where r is the number of relevant items in the top R items. The second metric is non-interpolated average precision rate (NAP) defined as:

$$NAP = \frac{1}{R}\sum_{i=1}^{R}\frac{i}{Rank_i}$$

where r is the number of relevant terms, and $Rank_i$ is the position of the i*th* relevant item in the ordering. Average precision rates at intervals are interpolated to avoid recall of zero.

In analyzing the effectiveness of the eigen-trend based methods for identifying research trends using the trec_eval tool, Tseng et al. found that a linear regression of the simple trend $D=[d_1, d_2, \ldots d_n]$ performed better than the eigen-trend methods [7]. Upon analyzing the data they found that the simple authority vectors which identify the authority of each source, is almost the same as the eigen-trend authority. Eigen vectors are most useful for feature reduction tasks. Chi et al. found that it performed well at identifying authority sources contributions. This is not

surprising as this is the type of task Eigen vectors are most useful for; isolating features and determining the direction of their vectors.

2.11 Burst Detection to Identify Burst Novelty

This dissertation was motivated by the observation that burst detection algorithms are inadequately equipped to access how novel bursty topics are. Therefore they do not effectively identify emerging trends in scientific literature. Tu et al. proposed a set of novelty indices to help mitigate this aspect of the burst detection problem [81]. The authors generated a novelty index (NI) and a published volume index (PVI) to identify the detection point (DP) of a topic. They define a topic's DP as the point at which the topic becomes emerging and valuable. For a novel topic to be considered emerging, its NI must be greater before the detection period than after, and, conversely, the PVI is higher after the detection period than before.

In this context, aging theory is used to model the life-cycle of a research topic to determine whether it is emerging and when it has stopped emerging. In the case of news the cycle modeled involved: birth, growth, decay and then death [82]. Applications based on this method use the concept of energy to indicate the different stages of the news cycle. The energy of a topic increases when a topic becomes popular, and decreases as its popularity wanes. Tu et al. adapted this method to the case of scientific research. To define NI it is necessary to first define the potential development year (PDY). The PDY is the first year to the current year when a topic will have no following years for which it has years with 0 publications (until topic death). This is the earliest period for which it can be seen to constantly growing in popularity. The NI is the inverse of PDY. If the PDY is 5, then NI is 1/5=0.20. To insure NI is between zero and 1 (without

normalization when NI and PDY are the same the result is undefined) it is normalized in the following way:

$$\frac{1}{k - PDY + 1}$$

where k is the current year. The PVI of a topic in the kth year is formulated as follows:

$$PVI_i(Topic) = \frac{SUM_i}{SUM_j}$$

where SUM_i is the accumulated number of papers from the first year to the ith year for the topic, and SUM_j is the accumulated number of papers for the set of articles analyzed. The DP is defined as the point at which NI and PVI intersect. It is the maximal value of the two indices of novelty and hotness. The detection point value is value of the DP when it intersects the Y-axis. If the Year of detection point (YDP) is i the VDP is calculated as follows:

$$\frac{PVI_{i-1} + PVI_i + NI_{i-1} + NI_i}{4}$$

To determine the effectiveness of their method, they analyzed title and abstracts from the ACM digital library. This is a large database of information systems and computer science journals. They validated their results by surveying previously published related works to find information about the candidate emergent topics, and interviewed experts on the topic. The first part of their experiment is to review how well their method detected emergent topics identified in the paper by Jo et al. [83], (who used a different dataset) to assess the effectiveness of their results. They analyzed the detection points of four topics u that paper and their selection of the year for which the topics could said to begin to emerge was the same as the results of Jo et al. They also interviewed five experts and asked them if they thought those areas were emerging during the year they identify as YDP for each, and their experts agreed.

There are several aspects of their methodology worth noting. First, using this methodology there is no way to assess the intensity level of a burst. They analyzed using known emerging topics, so they have not addressed the issue of noise in the data. This method could not be used for reemerging topics as they start by finding the first year for which there are no preceding zero counts for topic arrivals. They do not compare their method with the Kleinberg method, so it is difficult to know in what way their method is an improvement. Also, using expert review is problematic. As noted in Chapter 1 and section 2.2 it is very difficult for researchers to stay current of scientific trends. If that is true, how can we be sure an expert can pinpoint the time a topic is trending? It may be possible that the experts may recall the time an article was frequent, and not necessarily bursty. As mentioned in section 2.9 frequent topics are not always those that are bursty.

Other researchers such as Yin et al. [84] build upon Tu et al.'s methodology by using term relatedness methodology to generate a better candidate set of terms. Term relatedness may be a good way to reduce noise in the dataset.

2.12 Limitations of Existing Methods

After reviewing various burst detection methods, it becomes clear that one important problem is the lack of a clear set of requirements for burst detection methods. Consider Kleinberg's conceptualization of the problems. Kleinberg's model identifies more information about a burst than mere change in rate. By characterizing burstiness in terms of level of intensity, Kleinberg's method allows for an easy comparison of bursts. If multiple topics burst concurrently, their relative levels of intensity can be used to compare them. Also, this method characterizes burst duration at different levels of intensity. This method is flexible enough to characterize many

different burst structures. The model of He et al.'s addresses changes in rate but does not characterize bursts at different overlapping levels of intensity [57]. For some methods, e.g., topic drift, very specific assumptions are made about emerging topics that may make modeling many emerging trends with different citation structures difficult. Qian et al's topic drift based model assumes that top nodes actually always become the center of new clusters without demonstrating the validity of that assumption [76]. Recall may be a significant issue for their methodology as may be true for many of the citation-based methods as information about the relatedness of topics can be lost when text is not used. Tu et al. make assumptions about the growth model of topics [81], and the EM and eigen-vector based methods assume there are many parameters of unknown underlying features [55], [56], which does not seem to be an appropriate assumption in this space.

Another major limitation in the conceptualization of topic growth demonstrated by all the approaches is the assumption that bursts in the scientific literature should be modeled like bursts in the news. When a topic is popular in the news it means that many news programs mention the same topic. When a tweet is considered popular it means that the exact same words have been replicated numerous times. When it comes to scientific research, the popularity of a topic means that new discoveries have been made. In science, a topic remaining viable at a consistent publication rate or fluctuating in occurrence can still be associated with novelty and interest in the community. The lack of a consistent change in rate may merely be due to the difficulty of finding new discoveries. A topic burst, maintained at a given level, could be representative of an emerging topic. Methods such as Tu et al.'s, which assume a topic growth model like that of news topics, will miss many such emergent trends. Tu et al.'s method is designed to recognize only one burst from the time at which a topic has occurrences every year, without occurrence gaps. If there were occurrences of a topic every year between bursts ,the second burst would not be recognized.

Therefore, I consider the burst model to be best able to capture emerging terms as it does not depend on a specific growth model, but instead on intensity level.

Evaluation of methods is another significant problem in this space. Most of the work on burst detection, particularly involving detection of bursts in a stream of scientific documents, has defined evaluation metrics that are inappropriate for the task. The basic problem in that, as I have mentioned earlier, is they start with topics known to be bursty (e.g., the physics-based model, novelty indices method, the research front detection methods, and the EM method). These datasets have limited noise, and there is no indication on how well the algorithm would perform in an open domain discovery context. He el al. use MeSH, which does not allow for the discovery of new terms based on content in new documents in the document stream. Morinaga et al. generate topics with dynamic methods, which is better than many of the other methods. However, they only assign one topic to each document, which is limiting. If documents can only be identified with one topic, the frequency counts of other topics mentioned in the paper will be lower than they would if documents could have many associated topics. Also, Tu et al. use expert reviewers, despite the fact that they do not justify the assumption that an expert will be able to pinpoint the year a topic was bursty. As mentioned in Chapter 1 and section 2.2 researchers have a difficult time stay current on research trends, so this may not be a useful evaluation method.

In the next chapter, I introduce my framework for emerging topic detection. It relies on an overlapping hierarchy of topics, which classifies each document by a variable number of topics. It uses weightings based on the historical importance of documents to identify novel, though potentially reemerging bursts. The framework relies on characterization to of bursts, so that untested assumptions are not made. A fitness score is introduced to detect bursts early in the topic life cycle of bursts detected.

CHAPTER 3 A BURST DETECTION FRAMEWORK FOR REAL-TIME USE

3.1 Outline for Burst Detection Framework

I now propose a framework for identifying and vetting bursty topics in biomedicine that meets the specific needs of users identified in sections 2.1 and 2.2, based upon the Kleinberg burst detection model. This framework addresses many of the limitations of existing methods. For example, topic counts at each interval are weighted based on the historical importance of the topic. If a topic has experienced a high degree of activity for an extended period of time, its bursts should not be considered as important as those of new or reemerging trends. The main components of my model are as follows:

- Archival weighting: A novel set of weightings based on the historical importance of each topic identified.

- Fitness mode: Features such as journal impact factor, and associated funding data are used to develop a fitness score, for each of the topics to identity which new topics are likely to burst in the future.

- Bursts are characterized over an extended planning horizon for the cardiology discipline to understand what a typical burst trend looks like in this space and to better understand how to identify important or emerging trends.

- Cluster analysis is used to create an overlapping hierarchical structure of scientific literature at the discipline level. This allows for granularity adjustment (e.g. discipline level or research area level) in emerging topic detection for different users.

3.2 Archival Weighting

Research is cyclical. Ideas studied decades ago re-emerge. Research topics are ever expanding. The focus of the study of a particular disease changes over time as new developments are made related to treatment. Because of this tendency, it is possible for a new topic in a stream to appear to be a novel topic, without actually being new. For that reason I weight the number of publications at each interval. My weighting is based on the term frequency-inverse document frequency-weighting scheme (TF-IDF). TF-IDF is used to calculate the importance of a term for a given document based on how common the term for a collection of documents. If a term is common in the set of documents considered, the TF-IDF weighting would decrease its importance for each paper it appears in. This way, the terms that distinguish a particular document will be used to define it. TF is the frequency of a term in a document, and IDF is the log of the number of documents in the corpus divided by the number of documents where the term appears: tf * log (D/ (d € D: t € d)).

To weigh the number of times a term appears at each interval, inverse document frequency of the term for a previous period is considered. In a preliminary study of all cardiology journals represented in MEDLINE for the period of 2006-2009, IDF was calculated for the term for the period of 1995-2005. How important a topic was in the past, when determining its importance in the present, was considered. The metric developed, current term frequency inverse archive frequency (CTF-IAF), performed well in a preliminary analysis. The results demonstrated that our score did a better job of finding long-term bursting terms, with 63% of those terms continuing to burst in the following year. Using unmodified term frequency, only 25% of identified burst within the range studied continue to burst in the following year.

3.3 Granularity Adjustment

A burst of activity in this space can occur at many different levels of granularity. Burst detection models in the literature have mapped large-scale topic bursts like DNA, which have affected almost every biomedical domain. Bursts can occur within a smaller research area as well. In order to make burst detection *results* most useful to various interest groups, one must consider the degree of burstiness. To appropriately identify relevant levels of granularity, a topic hierarchy is used to create thresholds of topic frequency and document counts to categorize topics. A topic hierarchy provides the ability to further contextualize bursts so that a potential researcher could identify bursts similar to a set of topics.

3.4 Burst Characterization

In addition to weighting the arrival-rate of topics in a scientific document stream, characteristics of bursts were analyzed. Theories about the growth structure of topics have precipitated algorithmic design choices. Tu et al. assume that a topic's popularity in science is characterized by a birth, growth, decay and death structure. They then set out to identify trends by analyzing topic frequencies across time. They do not actually attempt to determine if this is the most appropriate assumption, as do the other approaches discussed in chapter 2.

Understanding what a typical burst structure actually looks like in this space can help to determine which bursts, at what level of intensity are of interest. For example, if one knew the typical burst duration for a topic at a given frequency range, one could set the planning window, or timeframe, to analyze with better accuracy. If one knew a burst typically lasts for 4 years, one would want to set the timeframe to identify bursts for a longer period. If one knew that a burst at a given level of intensity indicates that it is likely to stay bursty for at least 'x' years, that

information could be very useful in the attempt to find significant or important bursts. Further, if the assumption that research is indeed cyclical and there is ample evidence that a given topic can experience multiple bursts separated by a long time period, one can try to ensure that my algorithm can pick up on those bursts. Tu et al.'s burst structure would not allow for the identification for such bursts.

3.5 Fitness Score

When a topic is introduced into a particular field, it flourishes under certain conditions. There must be interest in the field, and researchers capable of dealing with it. The support it receives (in terms of related research that can back it up) is dependent on time, recent work, popularity, visibility of associated journals, and conceptual models of active researchers. Its likeliness to flourish, or its fitness, can be, to some extent, quantified and that can be used to predict its future growth. To try to identify topics that will be bursty at the earliest possible time, I develop my own fitness scoring.

There is a need for change within any given research area as researchers need to maintain a given threshold of productivity and/or establish themselves to be viable. Publication is dependent on new findings and methodology, shifts in focus, or the generation of new topics. Considering Small's model of scientific development one can understand why this is. Without change scientists can exhaust a research area, which then declines and may eventually become obsolete.

Citation growth is a process of accumulative advantage [58], [59], [85]. Citation networks demonstrate scale-free properties. The network grows based on preferential attachment, where some nodes are more likely to gain connections. Co-authorship networks demonstrate a power-law region followed by an exponential or Gaussian cutoff because of individual capacities to

collaborate[86]. A node's ability to attract new connections is not just based on preferential attachment, but on its fitness to connect to other nodes. Fitness, in this context, can be defined as in complex systems such as citation networks, nodes effectively compete for links in a constantly evolving system [87]. Competitive latecomers do have a chance to break up the network because of their ability to serve a particular purpose, as well as their ability to remain relevant as the context changes.

Networks evolve and the nature of that evolution can be derived from structural changes. Understanding that at a particular time the node introducing a change in a concept will alter the network makes it easier to identify those concepts. Researchers compete within a specific research area. In order for a research area to evolve, there must be new discoveries. Bibliometric research describing the evolution of research fields typically identifies conceptual changes in how researchers describe a subject area. However, there can be changes in approach and methodology that also give a competitive edge. Topics can drift from more specific topics to more general topics. Identifying shifts that will be associated with important bursts is necessary. A fitness score for a research topic can be generated by looking at various factors: early citation counts, impact factor of journals associated with it, and associated funding. Ke et al. [86] developed a fitness model to predict the impact of scholarly work at the document level which is relevant for my work:

$$k_i(t) \propto n_i \tau_i^{\beta}$$

where k_i is the node's degree over time, τ is the time factor, β is a scaling parameter estimated empirically, and represents a set of factors associated with the node's competitiveness. They identify several variables to associate with documents that can be used to predict the number of

citations they will accumulate. Their variables used to identify the fitness of a node are: merit of research presented in the paper, existing influence of authors and publication venue, and age of the paper. Their predictions of future citations counts were very close to actual values.

A model similar to Ke et al.'s fitness model, can be used to identification for important burst based on the extent to which it represents a fit mutation can be developed. To develop my model I consider the following:

- How quickly papers already associated with the topic gain citations. Papers with the same citations counts within a five-year period often have widely different long-term citation counts. A full citation history, (i.e. the number of citations received during each year), can be used to predict future citation counts as early citations appear to play a serious role in determining long-term impact[88].
- Journal impact factor will be another feature used for this fitness score. Average impact factor is among traditional scientific performance measures that help indicate how visible a document is [89]. It is calculated by dividing the number of citations a journal received in the previous two years over the number of articles published in that journal during the same period.
- Associated NIH funding is also a feature used for the fitness score for this work. Funding gives a sense of how impactful expert reviewers at funding agencies believe the work to be.

Each of the scientific archive measures, and the raw count measure with be weighted with a combination of weights from the selected features:

$$M\sum_{k}^{i}K_i$$

where M is the measure and K is the set of normalized features.

CHAPTER 4 ANALYZING CARDIOLOGY LITERATURE WITH THE BURST DETECTION FRAMEWORK

4.1 Characterizing Topics for Burst Detection

I used the STC and Lingo clustering algorithms [90] to generate an overlapping hierarchy to characterize topics in cardiology. Clustering algorithms were chosen over existing topic hierarchies such as MeSH and UMLS (Unified Medical Language System) for several reasons. According to the National Library of Medicine (NLM) the purpose of UMLS is to make it possible for researchers, clinicians etc. to create "conceptual connections to machine-readable biomedical information" [91], [92]. It is the largest collection of medical terms. All of the three tools that make up UMLS allow one to either manually map information to a controlled vocabulary (the way library of congress uses subject headings) or for purposes of text retrieval. Papers are tagged with MeSH terms once only, and those terms are never updated. UMLS has an advantage over MeSH in that it allows for the tagging of documents with terms irrespective of what terms would have seemed appropriate at the time of publication. The limitation of using UMLS in a real-world context where researchers are trying to stay up-to-date with current trends is that new emerging terms will not be included. For those reasons neither UMLS nor MeSH will not be used for the dissertation work.

Many of the clustering algorithms used for scientific emergent topic detection rely on citation-based methods to cluster documents. However, text based clustering solutions are more suitable in this problem space. First, as mentioned in section 2.7, articles tend to cite recent papers, so the relationship between papers is often lost. I believe this is one of the main reasons citation based methods have as yet not solved the problem of emerging topic detection to any significant degree. Also, emerging trends may be part of a large discipline-based trends and

researchers may only cited the research area specific papers that mention a given sub-topic even though their work is strongly related to other larger topics. To truly understand the relationship between new and old topics one must have a complete understanding of how they relate and text-based clustering methods will be useful in attempting to achieve this.

Clustering algorithms can be used on text collections to separate documents into meaningful collections. STC (suffix tree clustering) is an algorithm that uses a suffix tree to generate clusters. It keeps track of all n-grams of any length in a set of word strings while allowing strings to be inserted incrementally. A suffix tree is a data structure that essentially has a node for every possible phrase in a collection of documents. Clusters are either nodes with a high number of documents associated with it, or are a product of merged nodes. Labels are generated from phrases at each node. For the purposes of topic detection clusters need to be identified with terms that are the most representative of the set of documents and characterize the distinctiveness of these documents from other clusters. Most text-based clustering algorithms are not able to achieve this. For instance, methods like STC use linear algebra operations to compare texts. It is difficult to generate good labels based on a numerical comparison of similarity (43). Also, it is even more difficult to generate labels that are distinct with this method. Lingo is an algorithm that has been designed to help alleviate some of these problems [90], [94]. It is based on single value decomposition, and method that has been developed based on the assumption that words used in a similar context have similar meanings. This algorithm uses single-value decomposition to generate labels that for each cluster that help make it appear to be distinct from other clusters. The following two figures demonstration the difference between STC and Lingo. They show the clustering results for an ‘exome sequencing’ search of the PubMed dataset. This

network analysis was performed by me using the Carrot^2 toolkit[95] an open-source clustering application.

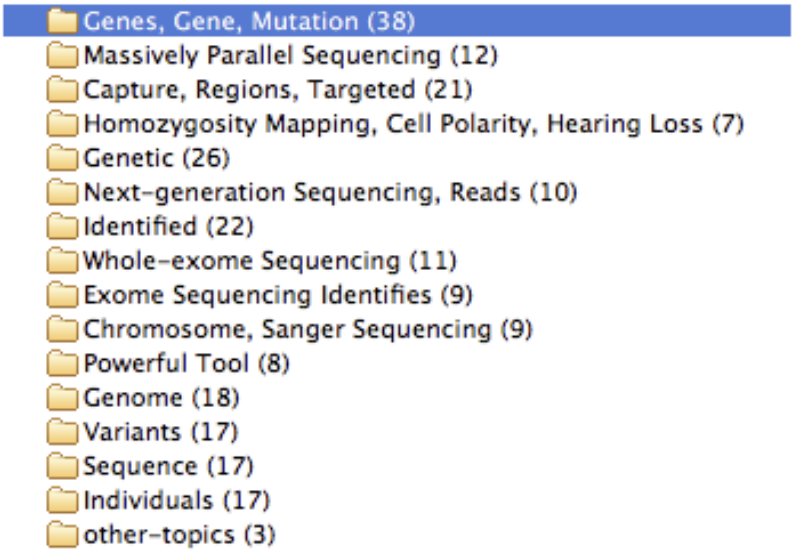
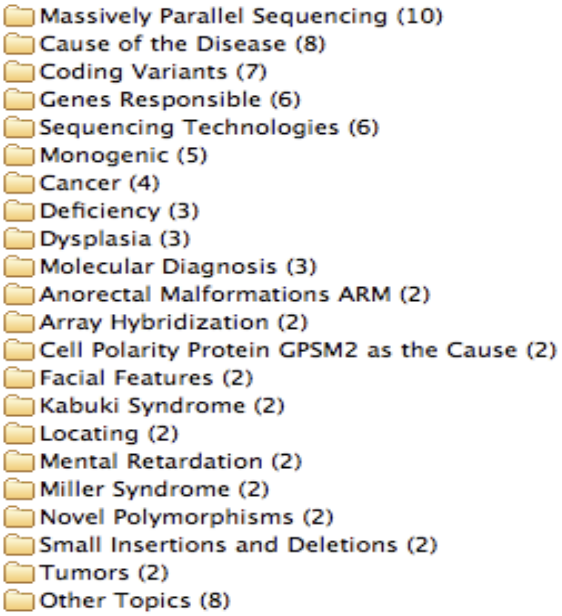


Figure 6 Search results for the STC algorithm.



|Figure 7 Search results for the Lingo clustering algorithm.

The STC labels are less specific than the results from Lingo, and also less distinct. In a comparison, by the creators of Lingo, of both techniques, using a cluster contamination measure on datasets previously partitioned by experts, Lingo produced purer clusters. The authors merged datasets with either very distinct or closely related data. The cluster contamination measure defines a cluster's contamination score as the number of pairs of objects in the same cluster, but not in any of the partitions, divided by the maximum potential number of such pairs in that cluster. The results indicated that the clusters identified by Lingo were less contaminated. In the analysis, Lingo also created significantly purer clusters than STC. Purity is a precision-based measure that focuses on the frequency of the most common category in each cluster. It is among the most widely used evaluation metrics for clustering methods [96].

One other aspect that makes the Lingo algorithm useful is that no assumption need be made about the number of clusters, a limitation of many other clustering methods [97].

Sub-clusters were generated and this process was performed recursively until cluster size is less than 10. For the period of 1990 through 2007, 195,441 documents were identified for clustering analysis. STC was used at the top level, as Lingo does not perform well, and clusters all documents together, when the document size is too large. Both Lingo and STC allow for overlapping hierarchies. At each level a document can appear in several clusters. At the highest level, manual inspection of output was performed. This was done because at the highest-level documents were clustered around general scientific terms such as significance, ratio, and analyze. Stop words, and stop label patterns were added to the existing list appear in the appendix. Due to the time-consuming nature of this task, the entire document set was not fully clustered. At the highest level 70% of the documents were clustered into meaningful clusters ranging from sizes of 61, 666 documents to 14. At the next level, and down to clusters of size 20

or less, 40% of total documents were clustered. These documents were taken from the top two clusters at the highest level: Left Ventricular, and Coronary Artery. Once cluster size was smaller than 10,000 Lingo was used to cluster documents.

4.2 Characterizing Bursty Topics

The discovery of bursts within a hierarchy of terms allows for the discovery of local bursts based on the area of interest for a given user. The following charts and tables, help characterize bursty behavior. All clusters with associated documents between 1990 and 1999 were analyzed using Klienberg's burst detection model. Statistical analysis was used to determine the relationship between cluster size during the timeframe and the volatility of yearly frequency. Bursts with less variability (or volatility) of their yearly frequency indicate bursts at a steady level during this period. Relative standard deviation is used to determine volatility of the trend-line. Levels of granularities are determined by analyzing the distribution of cluster sizes. For each level, bursts with the longest burst length, and burst strength are plotted on a yearly frequency graph. Topics experiencing reemerging bursts were also identified.

In finance, standard deviation is often used to measure volatility. This is accomplished by taking the standard deviation of the departures from the trade growth trend [98]. To better understand the volatility of bursty topics, relative standard deviation and linear regression was used to assess the behavior of topic trends. Relative standard deviation is standard deviation normalized by dividing it by the mean. The greater the relative standard deviation, the more volatile the trend. The correlation between relative standard deviation with: cluster size, the mean of yearly frequencies of the cluster, and the median of yearly frequencies is used to determine this relationship.

Table 1 Pearson's R correlation between relative standard deviation and three statistics.

Statistic	R
Total document count	-.083
Mean	-.121
Median	-.223

Table 1 shows that there is a negative correlation between relative standard deviation against the cluster size, the mean and the median. This indicates that bursty clusters with high frequencies are less volatile. Their trend-lines do not vary as much. Bursty clusters with low frequency have trend-lines that are more variable. This will have implications for burst detection, as it will be more beneficial to find the bursty clusters that are likely to stay bursty and not fluctuate in frequency too much. The following figures show the relationship between relative standard deviation and total count, mean, and median.

The following three tables further illustrate the relationship between relative standard deviation and the three statistics.

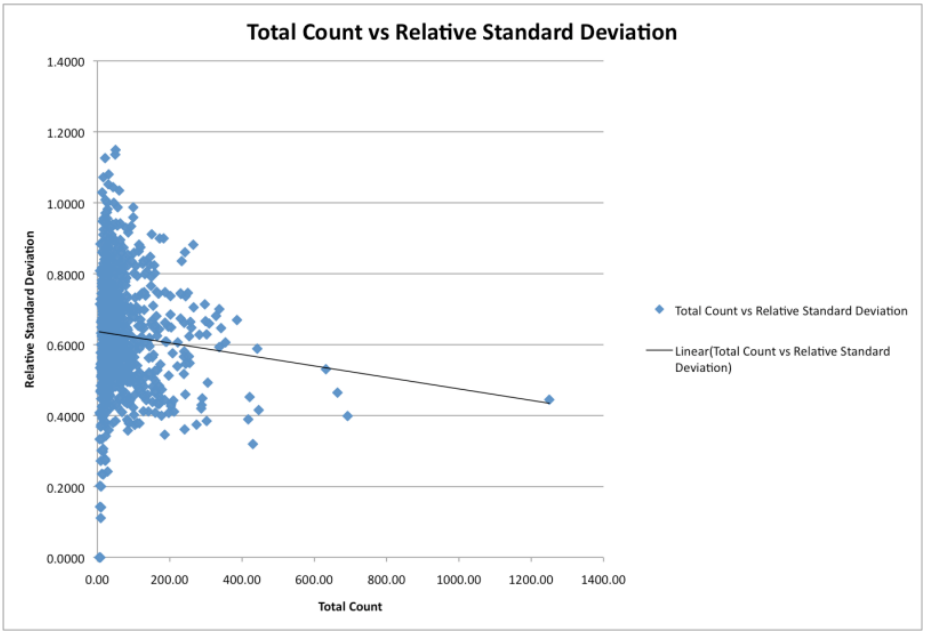


Figure 8 Total Count vs. relative standard deviation for bursty topics between 1900 and 2000.

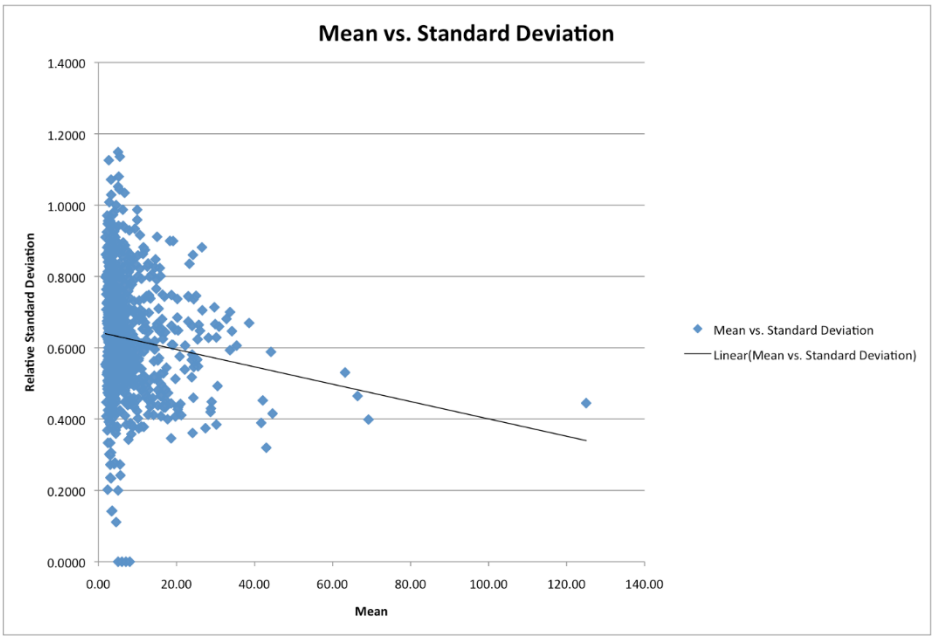


Figure 9 Mean for yearly frequency counts vs. relative standard deviation for bursty topics between 1900 and 2000.

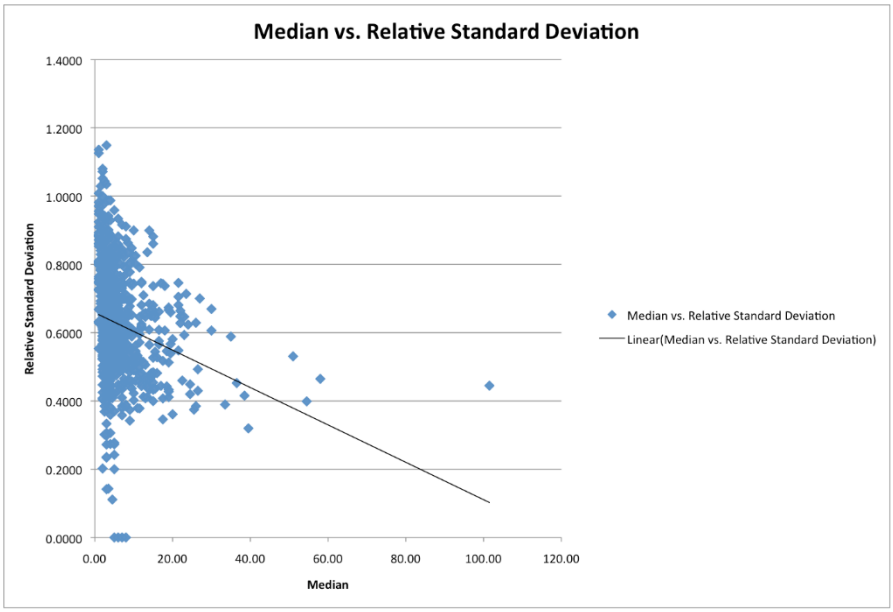


Figure 10 Total Count vs. relative standard deviation for bursty topics between 1900 and 2000.

Figures 8-10 show a negative linear trend. The trend for median of yearly frequency is the most pronounced. This is most likely because median is a measure more sensitive to outliers. In some cases, relative standard deviation is greater than one. This happens when standard deviation is greater than the mean due to many outliers. The graphs show that there are very

infrequent bursty topics that have low relative standard deviation, and do not have volatile trends. These are trends that typically do not have document counts in every year. At the low end of the scale there is a wide range of relative standard deviation scores. Because of the variability of relative standard deviation, investigation of different burst topic identification methods at different levels of granularity was done and is described in section 4.3.

Relative standard deviation gives a sense of the variation of frequency counts. It does not give an indication of the direction of the trend. It is possible for relative standard deviation to be high, even when the frequency counts are increasing steadily in a linear manner. Performing linear regression and identifying trend slope and r-square further highlights the relationship between total frequency and bursty behavior. Linear regression gives an indication of the direction of the trend. R-square is a statistical measure determining how well the data is to fitting the model. The higher r-sqaure is, the closer the data is to fitting the model. The set of bursts analyzed was limited to those that burst for at least three years total during the period of 1990 and 2000. There were 1038 bursts that meet those criteria.

Table 2 Correlations between r-square and three features for clusters with positive linear slope.

Statistic	R
Total Count	0.338
Mean	0.347
Median	0.328

Table 2 shows that the higher the frequency, the more stable the linear trend is. All three variables have positively correlated relative standard deviation--indicating that the higher the frequency of a topic the better it can be fitted to a linear trend. Understanding the trend of increase in frequency over time is crucial in developing a burst detection framework, as steady increases over time may be associated with more stable bursts.

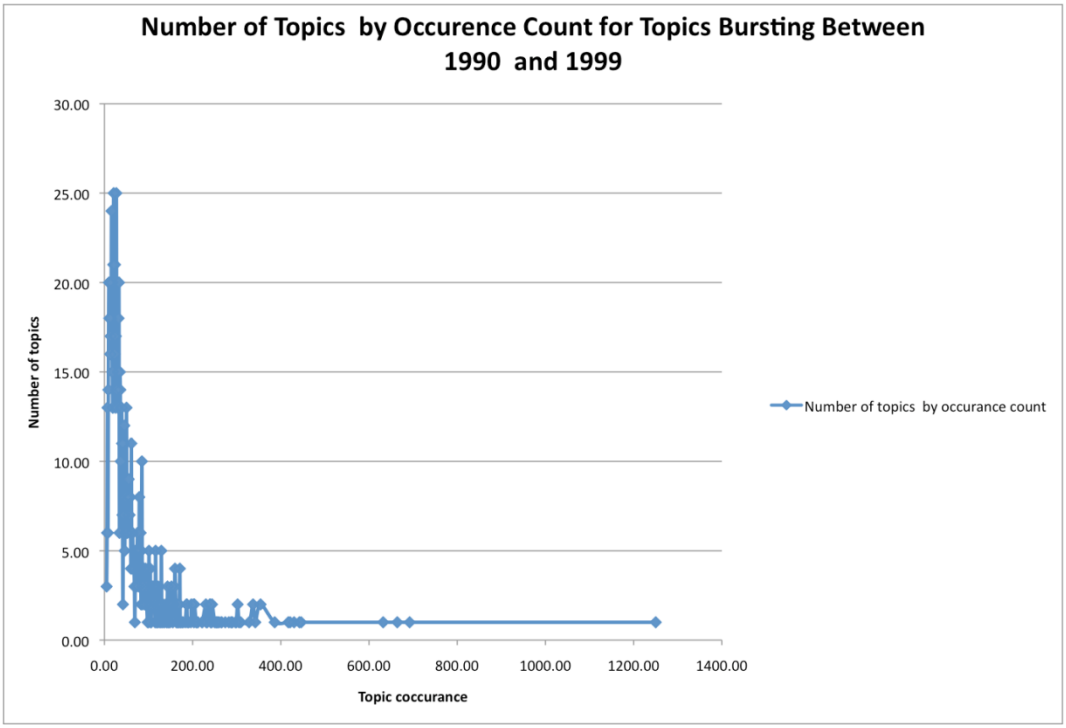


Figure 11 Number of topics by total count for the period between 1990 and 2000 that experience a burst of occurrence during that period.

The distribution of total count for bursty topics was examined to develop levels of granularity for analysis. The distribution in figure 11 is skewed to the left. Most topics, represented by clustered documents, have total frequency counts of less than 200, though there are a number of outliers, which have total counts greater than 400. The following figure shows the box plot for the distribution that is used to derive the levels of granularity for analysis of bursty topics.

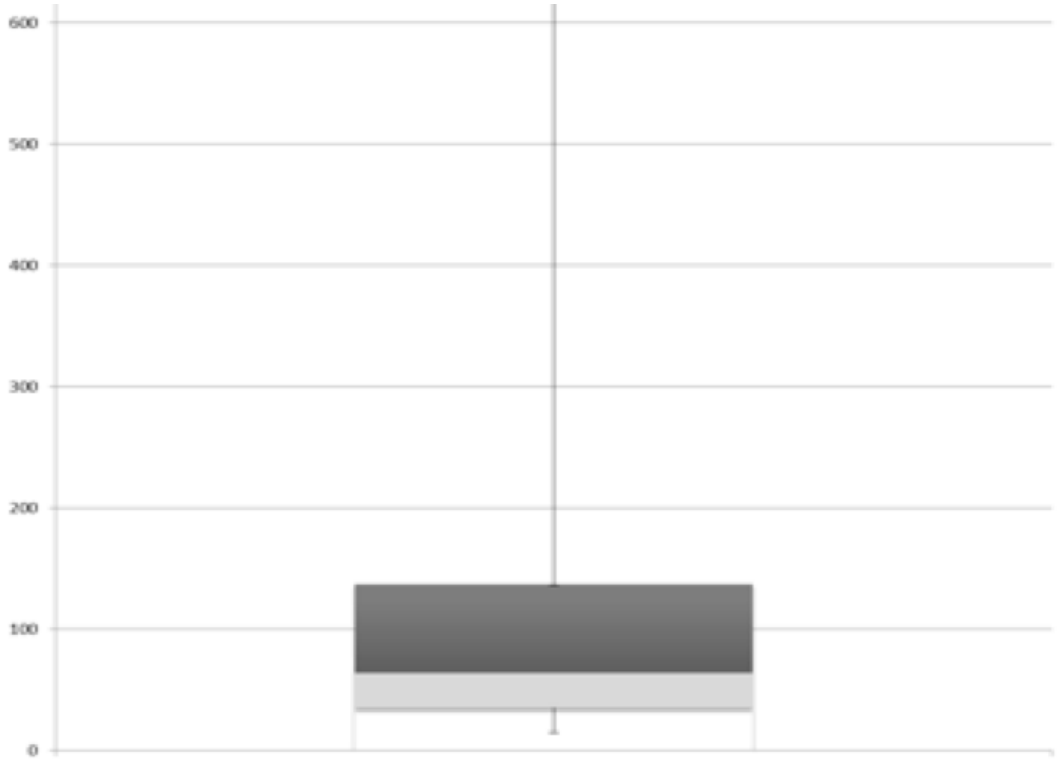


Figure 12 Box plot for total document count for bursty topics between 1990 and 2000.

Figure 12 shows that 75% of the clusters are of size 137 or smaller. The max cluster size (31,589) is not shown on the chart as it would make figure 12 difficult to read. The first quartile is 15-20, the section quartile is 20-35, and the third quartile is 35-137.. Because the distribution of the fourth quartile is so large, it was also separated into quartiles. Breakdown for forth quartile: the lowest 24% between are 138 and 169, the second 24% are between 169 and 222, the third 25% are between 222 and 342 and the last 25% are between 342 and 31,589.

The following figures display the trend-lines for top bursty topics in each range. Bursty topics were ranked by length and strength. Strength corresponds to the level of frequency during the period in which the topic bursts. The tables appearing below the figures indicate how many of the bursts occur at different burst lengths.

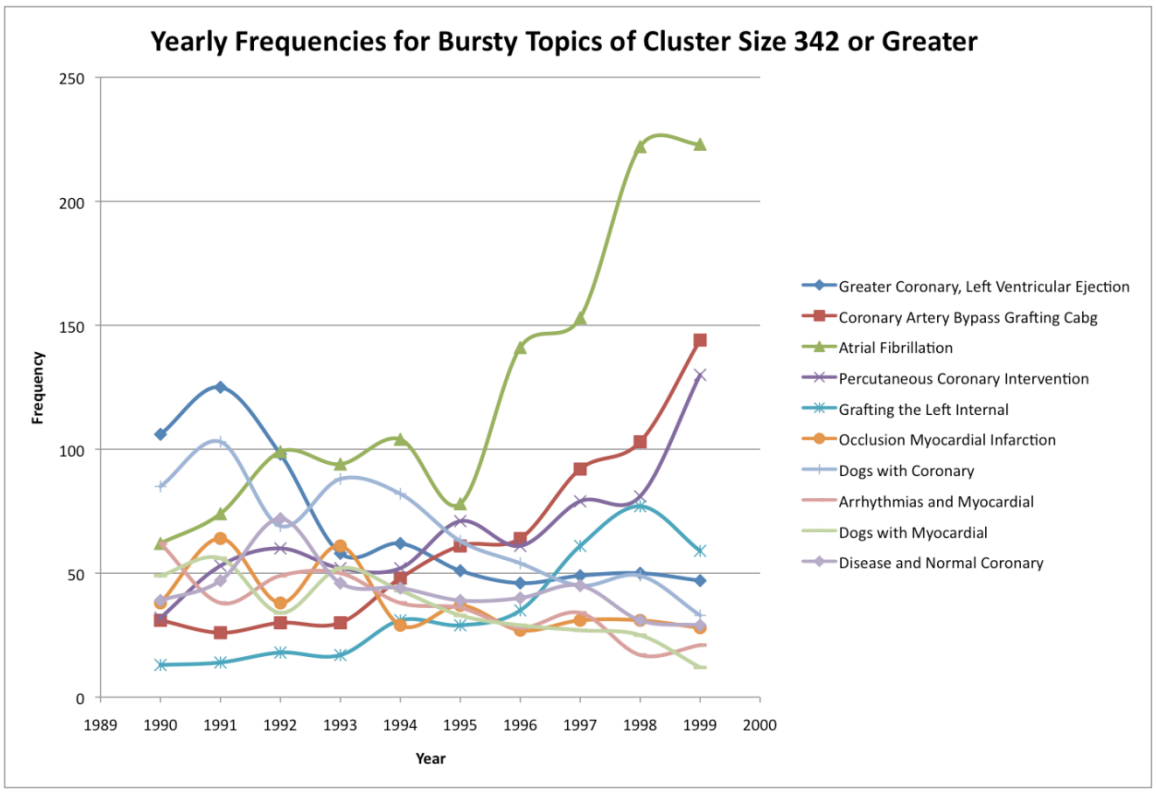


Figure 13 Yearly frequencies for bursty topics of cluster size 342 or greater. These are topics that have the longest burst length during this period.

Table 3 Number of bursts by length for topics of cluster size 342 or greater.

Length	Number of bursts
4	5 (10%)
3	7 (14%)
2	10 (20%)
1	28 (56%)

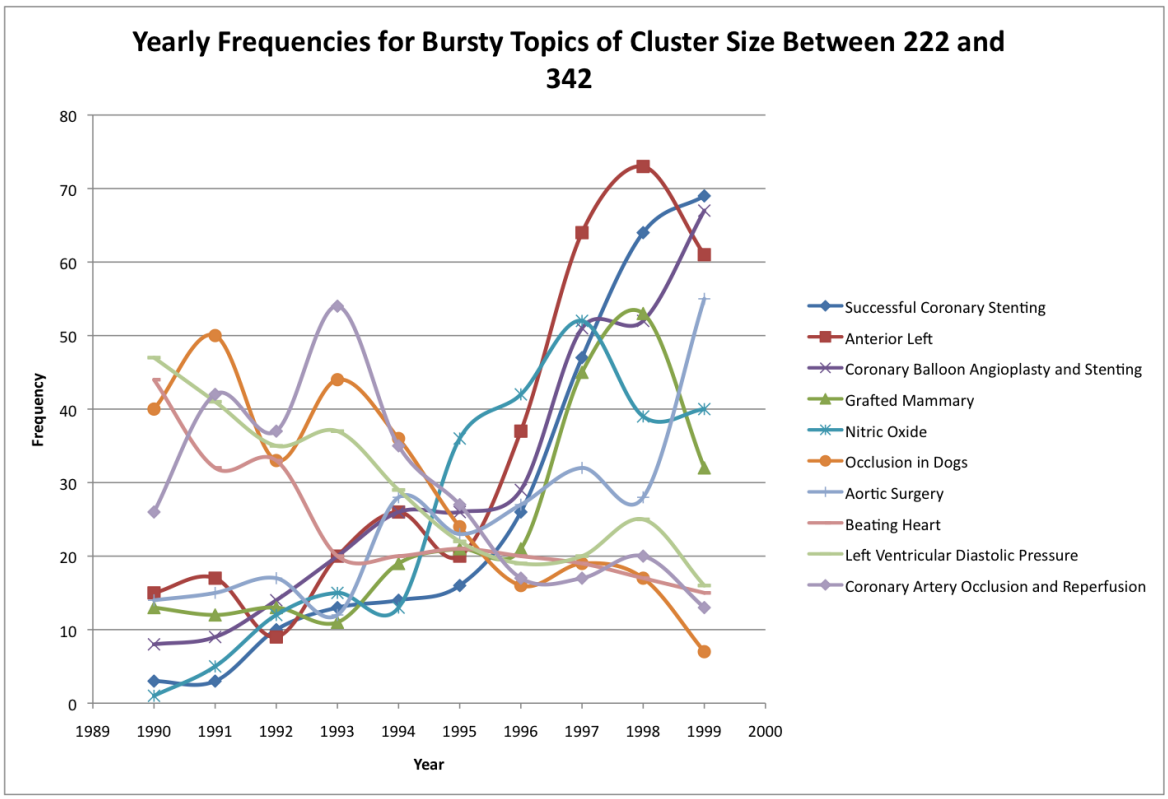


Figure 14 Yearly frequencies for bursty topics of cluster size between 222 and 342. These are topics that have the longest burst length during this period.

Table 4 Number of bursts by length for topics with cluster size between 222 and 342.

Length	Number of bursts
4	22 (19%)
3	14 (12%)
2	19 (17%)
1	58 (51%)

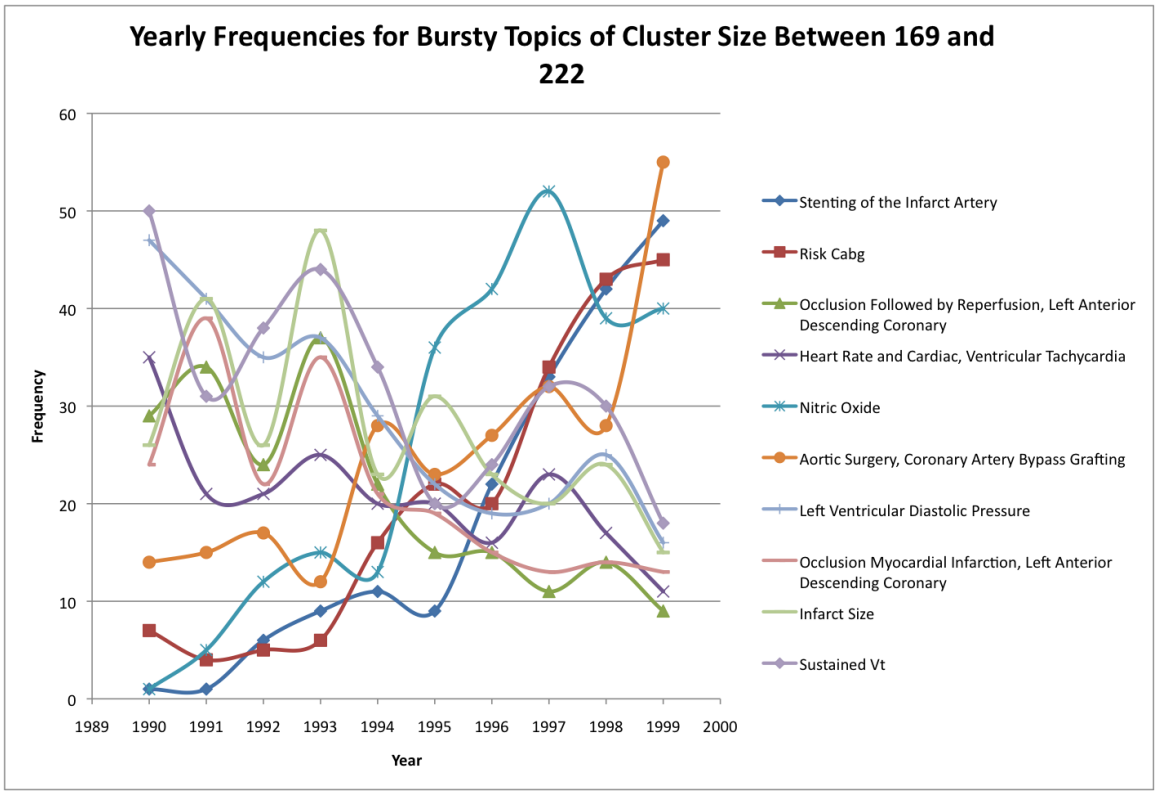


Figure 15 Yearly frequencies for bursty topics between cluster size between 169 and 222. These are topics that have the longest burst length during this period.

Table 5 Number of topics by burst length for topics with cluster size between 169 and 222.

Length	Number of bursts
4	12 (10%)
3	18 (15%)
2	25 (20%)
1	69 (56%)

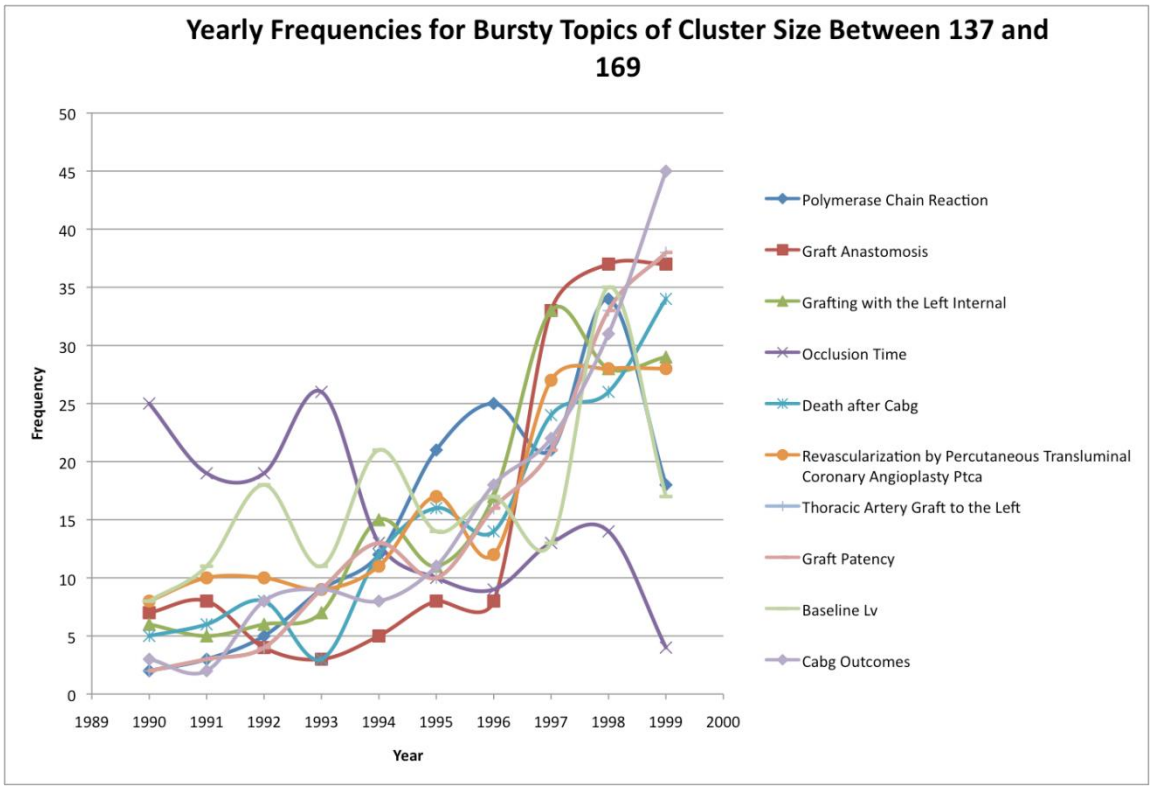


Figure 16 Yearly frequencies for bursty topics between cluster size between 137 and 169. These are topics that have the longest burst length during this period.

Table 6 Number of topics by burst length for topics with cluster size between 137 and 169.

Length	Number of bursts
5	4 (3%)
4	16 (12%)
3	16 (12%)
2	38 (29%)
1	56 (43%)

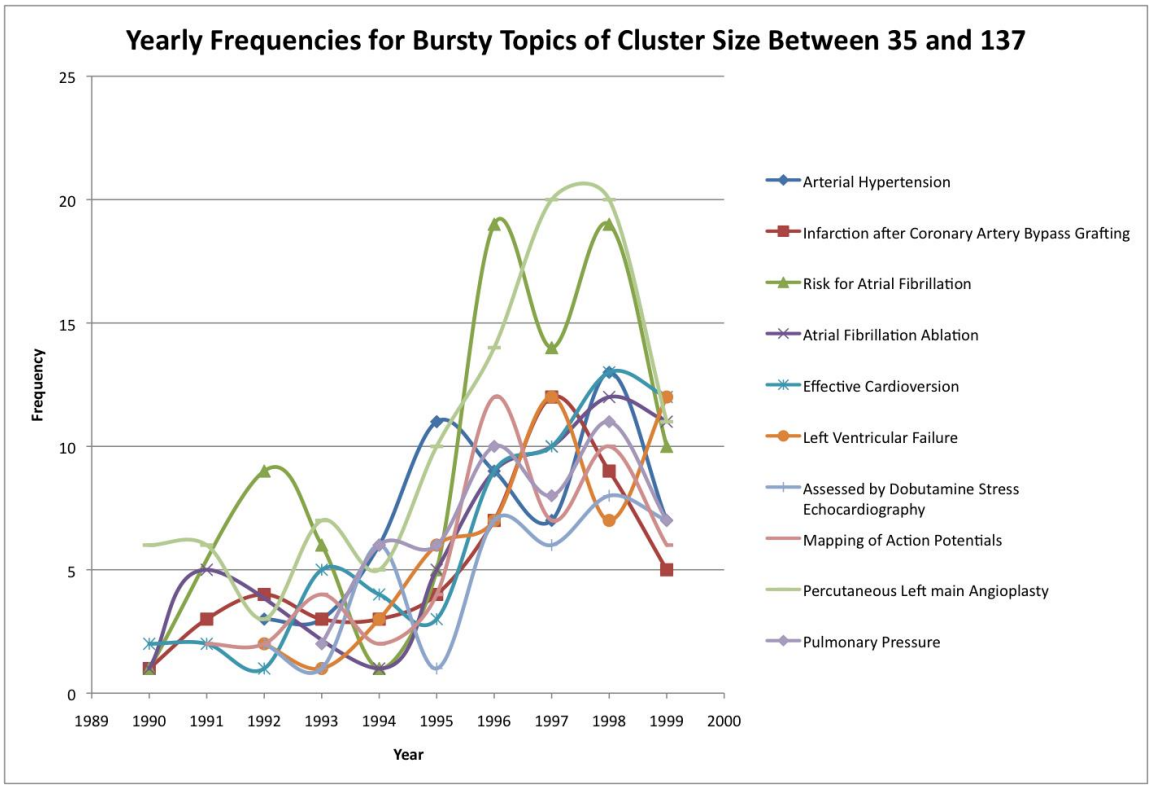


Figure 17 Yearly frequencies for bursty topics between cluster size between 35 and 137. These are topics that have the longest burst length during this period.

Table 7 Number of topics by burst length for topics with cluster size between 35 and 137.

Length	Number of bursts
7	1 (>1%)
6	2 (>1%)
5	28 (3%)
4	160 (14%)
3	252 (23%)
2	316 (28%)
1	361 (32%)

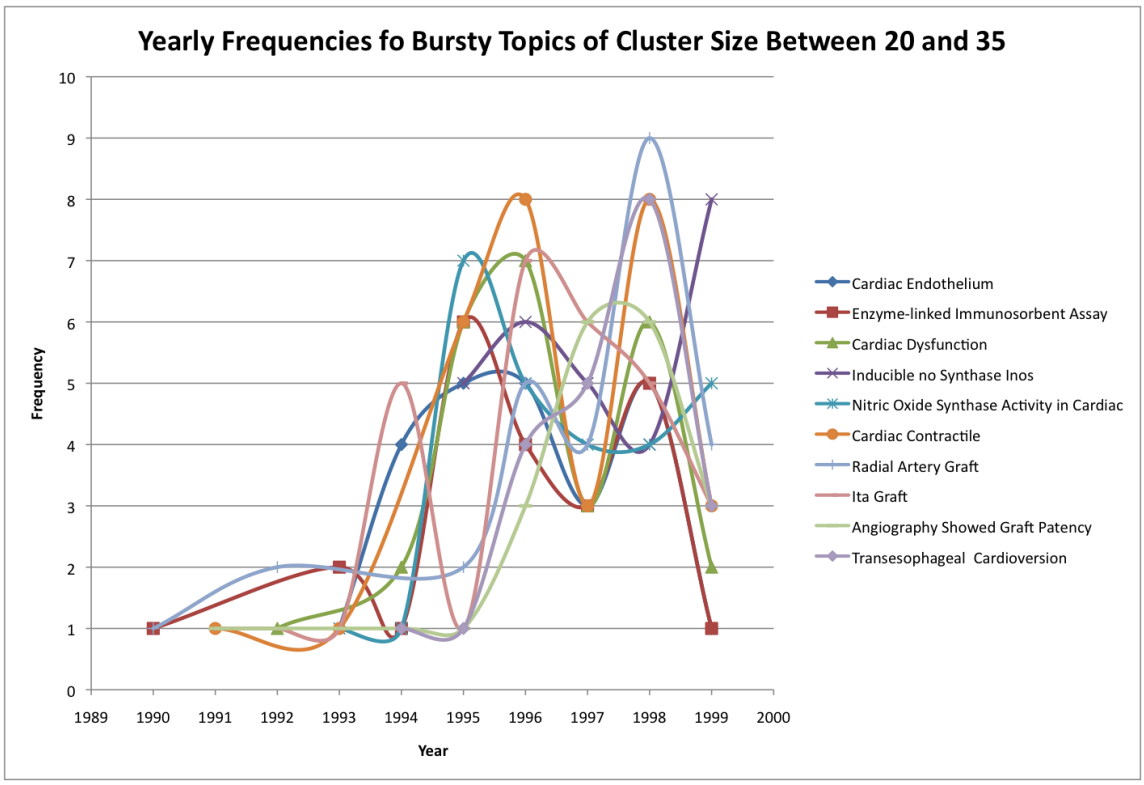


Figure 18 Yearly frequencies for bursty topics between cluster size between 20 and 35. These are topics that have the longest burst length during this period.

Table 8 Number of topics by burst length for topics with cluster size between 20 and 35.

Length	Number of bursts
7	2 (>1%)
6	12 (3%)
5	49 (11%)
4	69 (15%)
3	123 (26%)
2	129 (28%)
1	81 (17%)

Figures 13-18 demonstrate the differences in volatility at different levels of granularity. Figure 18 shows trends at the lowest level of granularity are unstable compared to the higher ranges. However, there is a significant burst for many of these terms with yearly frequency counts. For each of the ranges, with the exception of the range displayed in figure 13, the terms appear to fluctuate significantly. Terms with spikes in the 342+ range are terms that appear

frequently even in years where there are not spikes. Mane et al., who used burst detection to analyze PNAS data, were surprised that frequency and burstyness had a low correlation. The above charts help demonstrate the problem with Mane et al.'s assumptions about bursty topics. Highly frequent words are often already common. There may be a spike in the occurrence of the term for a few years, but it may not be as significant as it appears. Clusters that do have significant bursts do not become common quickly.

Tables 3-8 give an indication of the relationship between burst length and topic frequency. For topics in the top ranges burst length ranges from 4 to 1 years, with the majority bursting for two years at most. Though topics in the lower ranges appear to be more volatile, the burst length is often longer, with bursts ranging from1 to seven years.

Burst weight can also be used to characterize and understand bursts. The weight of the bursts is the reduction in cost of going from one state to another over its bursty interval. Words with high overall frequency have high weighting. Kleinberg proposed using weight to rank bursts [6]. The following table gives information on correlations of length and weight with mean, median and total count.

Table 9 Relationship between burst weight and various topic features.

Topic Feature	Burst characteristic	R
Mean	Length	-0.033
Mean	Weight	0.664
Total Count	Weight	0.650
Total Count	Length	-0.024
Weight	Length	-0.003

Table 9 shows that weight is negatively correlated with length and appears to be a function of total frequency. This indicates that weight is not a good criterion for ranking, in this context, as Kleinberg suggested.

The following chart displays the trends for terms selected from the above charts that demonstrate reemergence.

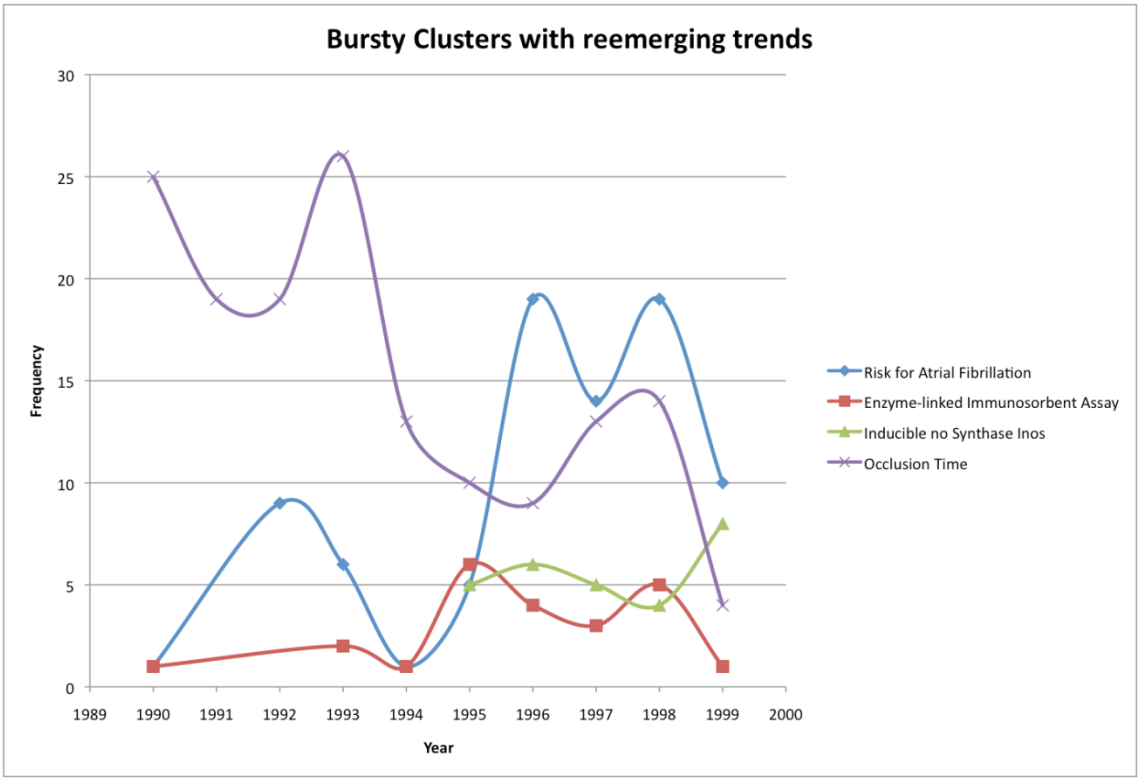


Figure 19 Results for terms whose frequency demonstrates reemerging burstiness.

The cyclical behavior of the trend-lines displayed in Figure 19 indicates that emerging trends can indeed be associated with old topics. In fact, during the period analyzed, 75 of the topics that burst have two separate topic bursts. Emerging trend detection algorithms will be very limited if, as Tu et. al. assume, topics burst once only. Topics that experience reemergence can be associated with new developments. Researchers whose work is related may be very interested in those reemerging trends.

4.3 Weighting Topic Frequency Counts with Archival Measures

To determine how best to account for the scientific literature archive, three weights were developed based on the historical frequency of each topic. These archival measures were then compared them with un-weighted burst detection results. The first measure is CTF-IAF, or current term frequency-inverse archive frequency:

$$ctf \times (\log(N7yr/7yrtf))$$

where CTF is the number of documents using the term in the current window of interest, N7yr is the number of documents in the seven year period which represents the archive, and 7TF is the number of documents in the cluster for the seven year period. This is the simplest measure I use to account for the cluster history.

The second measure is CTF-SQRT-IAF or current term frequency -square root- inverse archive frequency:

$$ctf \times (\log\sqrt{N7yr}/7yrtf))$$

The above equation is almost identical to the first one with the exception that instead of using N7yr it uses the square root of N7yr. This measure penalizes terms that appear frequently even more than CTF-IAF. Taking a log of smaller number will result in even lower values. Using this measure, terms that appear frequently in the archive, will have CTF-SQRT-IAF close to 0 or at 0 even if they appear frequently currently. This is an attempt to find bursts of terms that appear very infrequently in the archive.

The last measure is CITF-IAF, which stands for current inverse term frequency-inverse archive frequency:

$$(\log(cn/ctf)) \times (\log(N7yrtf/7tf))$$

This last one differs from the first in that instead of TF at the beginning there is log(cn/ctf) which is the log of the current document count divided by the current term frequency. This accounts for the prominence of a term in the current window of interest. I compare these measures to TF, which is the raw count of each term in the window of interest.

As discussed previously, the arrival rate for science must be discretized as journals publish in consistent intervals. I chose a window of one year. This is the smallest period of time in which activity can be seen for most journals. The planning horizon for this project was a ten-year period. The period of 1997 to 2006 inclusive was chosen to identify bursts. The percentage of bursts in that period for which there was a burst in the following year were identified. The results are displayed below. Kleinberg's method implemented in the Network Workbench Toolkit[2] was used to find bursts.

To understand how infrequent bursty clusters identified by each measure were in the 7yr historical period are, IAF scores were plotted for each burst against the number of clusters of that IAF score for each measure. IAF scores are large for clusters with low membership in the previous period. For this study I identified a planning horizon of bursts between 1997 and 2006 inclusive. The window size was one year. To generate IAF scores, cluster membership counts were generated for the period between 1990 and 1996 inclusive. IAF scores were rounded to the

[2] http://nwb.cns.iu.edu/

first decimal place. Each measure was used to identify bursts. The results are given in the figures and tables below. Figure 20 displays the number topic burst counts by IAF scores for each measure.

Figure 20 Displays the number of concept burst by IAF score for the non-weighted topic count.

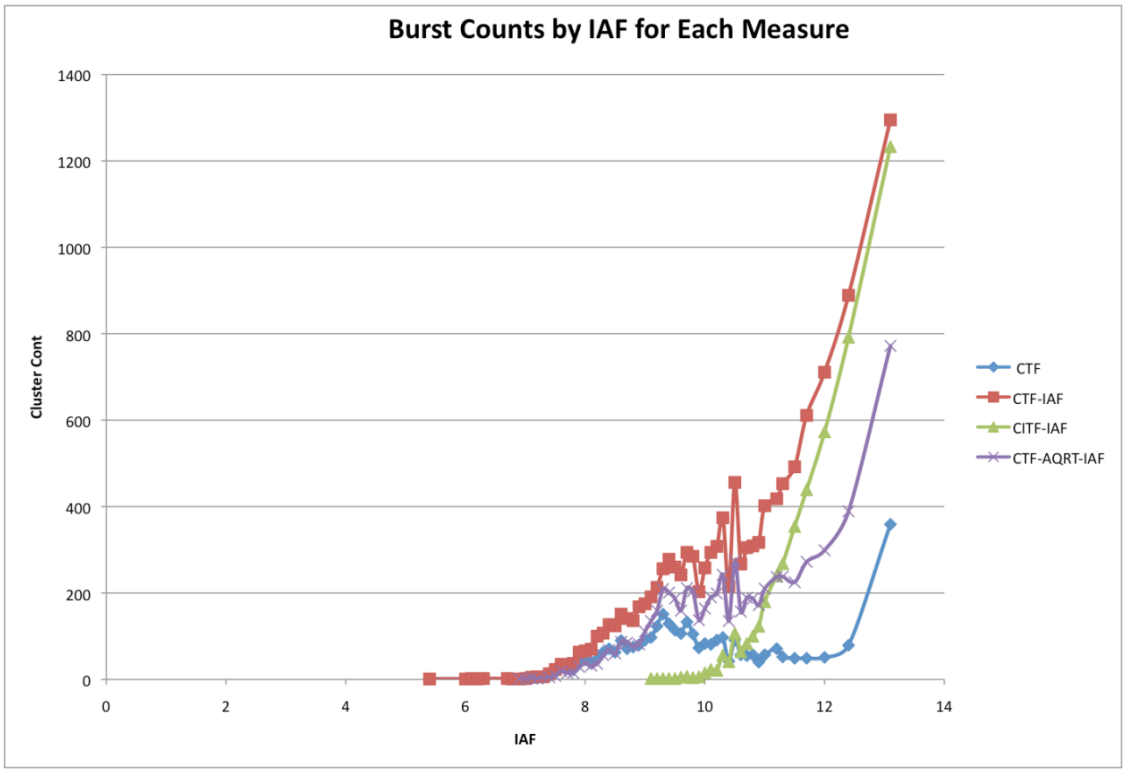


Figure 20 demonstrates a noticeable difference in IAF ranges, which produce bursts for the different measures. CTF, CTF-IAF, and CTF-SQRT-IAF, all have similar shapes, but have different rates of increase. CITF-IAF has the smallest range of IAF scores. It only identifies as bursty, topics that are historically rare. Not only are the ranges different the number of bursts found at each IAF score differs. This tells us that the some scores are better for finding bursts for very rare scores.

A methodology was developed to reduce the number of bursts analyzed. Each of the measures identifies bursts that have an overall low frequency of occurrence--especially the

weighted measures. Manually sifting through these bursts would be incredibly time consuming. At each level of granularity there will be bursts that are associated with high burst strength, significant frequency, high average impact factor and a high-normalized impact factor. These features will be used to select the top burst by measure. Also, based on an analysis of typical burst behavior at each level of granularity a threshold of frequency will be set to reduce the set of bursts found.

For each measure correlations between burst strength in the first part of the time period, and future burst strength were determined. Noise reduction in this step is based on cluster size. The cluster membership during the period of analysis must be 30 or higher, and the length of bursts must be 3 years or more. The following four charts show the cluster size distribution for bursty topics for each measure.

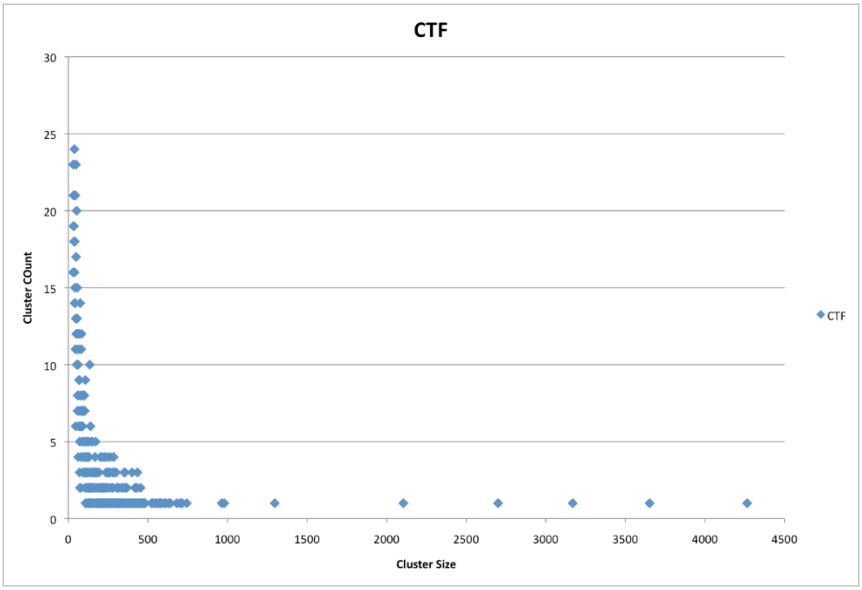


Figure 21 Cluster count by size for topics bursting at least three months with cluster size 30+ for un-weighted topic counts.

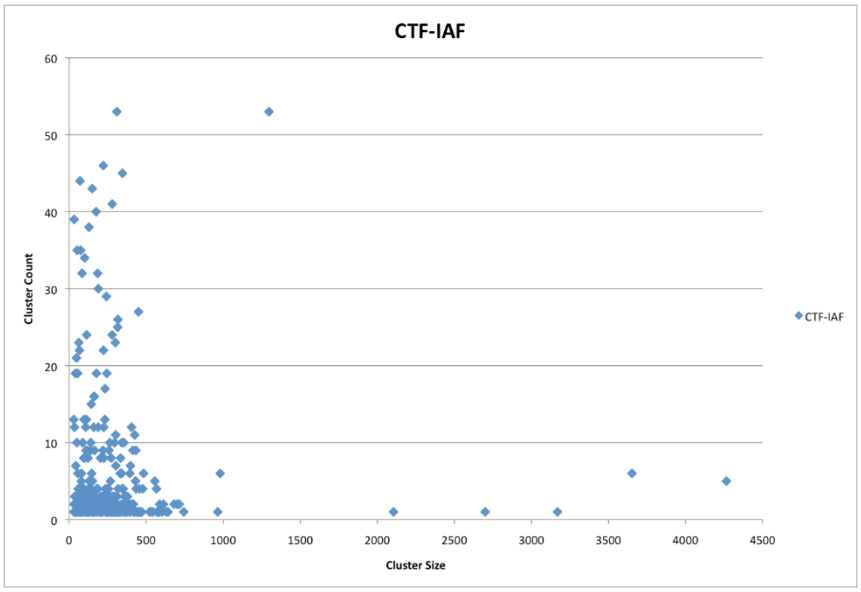


Figure 22 Cluster count by size for topics bursting at least three months with cluster size 30+ for clusters frequencies weighted with CTF-IAF.

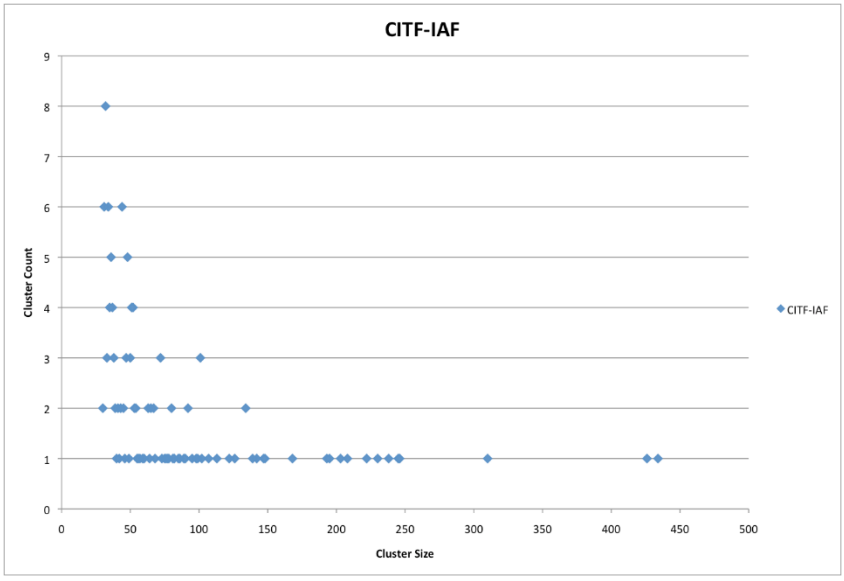


Figure 23 Cluster count by size for topics bursting at least three months with cluster size 30+ for clusters frequencies weighted with CITF-IAF.

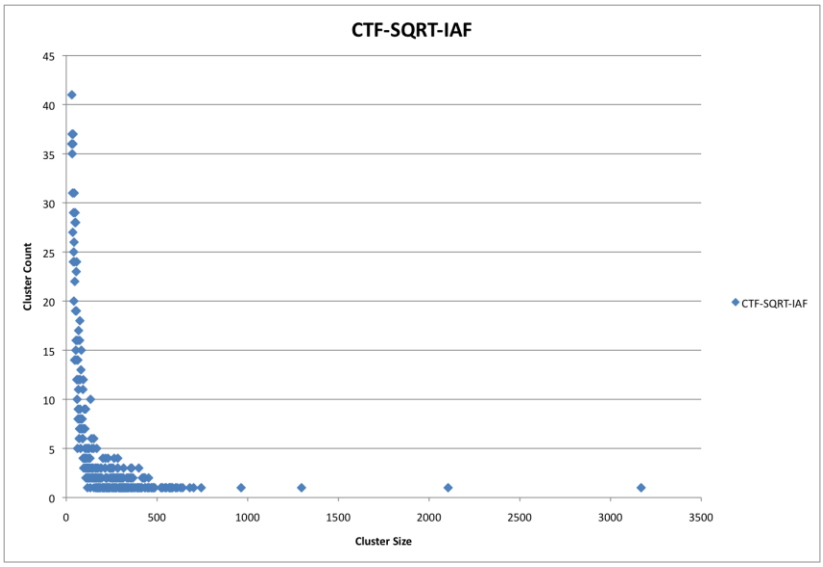


Figure 24 Cluster count by size for topics bursting at least three months with cluster size 30+ for clusters frequencies weighted with CTF-SQRT-IAF.

Figures 21-24 show the distribution of burst count is similar for each measure except for CITF-IAF, which has a smaller distribution range. Each distribution is skewed toward the left. With the exception of CITF-IAF, the distribution is similar to the distribution used in section 4.2. Since these distributions are so similar, the levels of topic granularity used in section 4.2 those levels will be used in this section.

To make a burst detection framework useful, it is important to identify bursty topics that remain bursty. The following table shows the percentage of topics that burst between 1997 and 2003 and remain bursty between 2003 and 2006.

Table 10 Results by weighted frequency showing the percent of bursts found which burst in the future.

Measure	Total number of bursts	Total number of bursts between 1997-2003	Total number of early bursts still bursting after 2003	Percent still bursting after 2003
CTF	1218	957	510	53%
CTF-IAF	1958	1563	735	47%
CITF-IAF	143	143	113	79%
CTF-SQRT-IAF	1628	1340	498	37%

Table 10 shows the differences between the measures in terms of their total number of bursts identified, and their ability to identify long-lasting bursts. This research indicates that when trying to identify how new a term truly is for a given area, one can increase the precision of burst detection methods. In this sense, precision would correspond to bursts which remain emergent. Precision is higher for CITF-IAF than any other measure. In terms of trying to find lasting bursts, it identifies the fewest false postivies. CTIF-IAF and CTF-SQRT-IAF identified the most bursts, though many of them are not long lasting.

Another important feature of a useful burst detection framework is the ability to identify bursts earlier than other methods. The following table shows which measures find bursts earlier than other measures.

Table 11 Displays the number of topic bursts a particular weighting identified earlier than another weighting.

Measure	Earlier than CTF	Earlier than CTF-IAF	Earlier than CITF-IAF	Earlier than CTF-SQRT-IAF
CTF		3	0	9
CTF-IAF	138		0	105
CITF-IAF	47	53		49
CTF-SQRT-IAF	132	108	0	

Table 11 demonstrates that weighted topics are better able to detect bursts identified by raw counts than the reverse. Also, no measure is able to detect bursts earlier than CITF-IAF. This indicates that it does a better job of early burst detection, when compared to other measures.

Linear regression on the current timeframe was performed, for each bursty topic identified by each measure that experiences a burst of at least three years total, was performed. This was done for topics with bursts that either ended in 2006, or were still bursting at the end of the period.

Table 12 Average R-square, total number of bursts, and average relative standard deviation for each measure.

Measure	Average R-square	Total number of bursts	Average Relative standard deviation
CTF	.46	783	.62
CTF-IAF	.39	1197	.59
CITF-IAF	.41	143	.56
CTF-SQRT-IAF	.48	883	.59

Table 12 shows that CTF-SQRT-IAF has the highest r-square, while CTF has the highest relative standard deviation for the period. High r-square means the data fits the liniear trend the best. A measure that finds trends easy to model is very important, as it can find trends easier to predict. To improve the results of other methods a fitness score will be used so that the most stable as well as newly bursting topics can be found.

Burst slope identification was done to assess how well each measure detects topics that have a significant increasing trend. This was achieved by creating burst strength trend-lines at each level of granularity. Linear regression was performed to find the slope of each trend line so that increases in intensity can be assessed, and compared. Only bursts with positive slope were used in this part of the analysis.

Table 13 Average slope by range for each measure.

Range	CTF-IAF	CITF-IAF	SQRT-IAF	CTF
Avg. for cluster size 342+	5.50	2.00	5.70	5.90
Avg. for cluster size between 222 and 342	3.36	2.9	3.27	3.50
Avg. for cluster size between 169 and 222	2.95	0.9	2.56	2.97
Avg. for cluster size between 137 and 169	2.57	3.60	2.81	2.88
Avg. for cluster size between 30 and 137	1.45	2.53	1.71	1.90

Table 13 shows that at the highest level of granularity, CTF finds bursts with higher average slope than the other methods. This indicates that at those levels it is identifying bursty topics with high average yearly increase in frequency. At the lower levels of granularity, CITF-IAF finds bursts with higher average slope than the other methods. This indicates that at those levels it is identifying bursty topics with high average yearly increase in frequency.

Google scholar data was used estimate the degree to which the documents identified by each burst are cited in the future. This will give an indication of the impact of the field and its continued stability. Some identified bursts may not continue to burst, but begin to remain constant. However, if the documents from the bursty period have on average higher citation counts than they had previously, the identified burst can be deemed valid. For each cluster, the

sum of citations counts were be analyzed for each range. The average of that number for each measure is displayed in the following table.

Table 14 Average citations for bursty clusters by measure.

Range	CTF-IAF	CITF-IAF	SQRT-IAF	CTF
Avg. for cluster size 342+	2101	N/A	1545	2062
Avg. for cluster size between 222 and 342	956	1402	944	995
Avg. for cluster size between 169 and 222	559	1155	566	627
Avg. for cluster size between 137 and 169	437	1113	400	443
Avg. for cluster size between 30 and 137	182	310	173	200

In the highest range CTF-IAF has the highest average citation count. For each other measure CTF has a higher average citation count than CTF-IAF. For each range except the highest one CITF-IAF consistently has the highest average citation count. This indicates that at the highest range CTF-IAF identifies bursty topics with high impact, and at all other ranges CITF-IAF identifies bursty topics with high impact.

4.4 Early Burst Detection with a Fitness Model

The results from section 4.3 demonstrate that my framework improves upon existing methods. To increase the effectiveness of this framework, a fitness score was developed. A

fitness score can help identify bursty behavior early in a topic's life-cycle. Early signs of prominence could help researchers identify an impactful research agenda. In the following subsections I identify my weightings for my fitness model. Once bursts have been identified for each measure. The effects of each fitness feature are examined. It will be important for future analysis to pinpoint which feature improves which measure, and to what extent. Normalized journal impact factor and normalized h-index are both impact measures. Funding data gives an indication of how important a topic is to policy makers. It would be useful to know if all features are necessary for improving bursts, as citation data is not easy to come by. This will also indicate which features should be weighted more.

4.4.1 Weighting with Normalized H-Index

Early impact in terms of citations is a good sign that a topic will continue to be a focus of research. H-index is citation-based method to determine impact of documents. A normalized h-index measure was developed to weight each measure from section 4.3. A researcher has index h if they have h papers with at least h citations [99]. It is a useful measure because it combines productivity with impact as it relates to not only how many papers published, but also how often they are cited. It is additionally not sensitive to extreme values and hard to inflate [100].

However, for the timeframe of my initial analysis h-index could not be generated, as Google scholar did not have information on citation for the overwhelming majority of papers in 1997 and 1998. Because H-Index is a good fitness measure, I performed a different analysis for H-index. To determine what would be a reasonable timeframe to expect enough citations for a set of documents to reach a relatively high h-index I analyzed the citation half-life data provided by JCR. The citation half-life of a document is the number of years after publication in which it

receives half of the citations it will ever receive. The average citation half-life is made available for the journals in this study by JCR. For cardiology the average citation half-life is 4.9. The maximum is 9.9 and the minimum is 1.4. If most of the citations are in a short time frame then citation counts for documents in the previous few years can be used to bolster occurrences of the term in the current time frame.

Normalized h-index score was developed in the following manner. First, of interest is not simply the impact of topics, but the impact in terms of the total number of documents. Also, comparing impact across topics even with the topic count is different is necessary. To achieve that the following formula h/d was developed, where d is the total number of documents associated with a given term and h is the h-index. This will allow for a simple weight with values between 0 and 1. The closer normalized h-index is to 1, the more documents were cited at least d times. The closer normalized h-index is to 0, fewer documents were cited at least d times. An analysis of h-index data using the data generated by my clustering analysis was not possible. During the period of 1997 through 1999 very little citation data was available. Because h-index has potential as a fitness measure I include the results from a preliminary study about h-index.

The setup for this experiment was as follows. The interval for document arrivals is six months. The time frame for the experiment is January 2004 to December 2009. Each documents was tagged with UMLS terms. For each term, normalized h-index was be calculated based on citations of documents published between January 2004 and December 2006. The reason citations are only gathered for 2004 to 2006 is that by the start of January 2010 they will receive most of the citations they will ever receive, and is therefore a good reflection of their scientific impact. Google scholar was web-crawled for citations for all cardiology documents for the period of 2004 to 2006. For each term normalized h-index was calculated based on the results.

Each measure discussed in section 5.1 was multiplied by normalized h-index and the results are displayed in the table below.

Table 15 Early detection results for measured weighted with normalized h-index.

Results	ctf-iaf * (1+nh-index)	citf-iaf *(1+ nh-index)	ctf-sqrt-iaf*(1+nh-index)	ctf*(1+nh-index)
Bursts detected earlier than unmodified measure	6406 (22%)	4123 (8%)	2700 (33%)	997 (41%)

Table 15 demonstrates that weighting with normalized h-index make it possible to detect some bursts at an early time. However, for the reason mentioned earlier in this section, h-index is not used for the complete fitness score.

4.4.2 Weighting with Normalized Journal Impact Factor

Average journal impact factor was used for terms for each time interval to weight each measure. Increases in journal impact measure can help bolster the apparent burstiness of terms. Journal impact factors were obtained from the Journal Citation Reports. I analyzed impact factor for journals in cardiology to determine the best way to normalize impact factor. First, I determined the maximum average impact factor for clusters for each year in this area. Then I determined the average impact factor for each year. To normalize impact factor the average journal impact factor for each cluster is determined. The normalized funding measure is:

for clust-avg>year_avg

Norm-impact= (clust_avg-year_avg)/year_max-year_avg)

Each weighted measures was multiplied by 1+ Norm-impact. To determine whether journal impact actor is in fact a good criterion to weight term counts with the following experiment was performed. The time frame for the experiment is the same from the previous section. For each cluster the normalized journal impact factor was calculated and each measures was multiplied by normalized impact factor.

Table 16 Early detection results for measures weighted with normalized impact factor.

Results	ctf-iaf * N-Impact Factor	citf-iaf * N-Inpact Factor	ctf-sqrt-iaf*N-Impact Factor	ctf*N-Impact Factor
Bursts detected earlier than unmodified measure	348 (3%)	2(>1%)	193(3%)	121(10%)

The results from table 16 demonstrate that weighting with normalized impact-factor make it possible to detect some bursts at an early time. Weighting with impact-factor has the greatest affect on the un-weighted measure.

4.4.3 Weighting with Normalized Funding

NIH-Reporter funding data is also used for the fitness measure[3]. Funding associated with each pmid in each cluster by year is averaged. The average rate for clusters for each year is determined, as well as the maximum average. The normalized funding measure is:

[3] http://projectreporter.nih.gov/reporter.cfm

for clust-avg>year_avg

norm-funding=(clust_avg-year_avg)/year_max-year_avg)

Each weighted measures was multiplied by 1+ Norm-funding.

Table 17 Early detection results for measures weighted with normalized funding.

Results	ctf-iaf * N-funding	citf-iaf * N-funding	ctf-sqrt-iaf*N-funding	ctf*N-funding
Bursts detected earlier than unmodified measure	1097 (9%)	1(>1%)	629(9%)	469(14%)

Weighting each measure with funding data finds more bursts earlier than weighting with impact factor. Similarly to impact factor, weighting with funding has the greatest impact on the un-weighted measure.

4.4.4 Weighting with Combined Features

Normalized funding, and journal impact factor were combined to create a more robust fitness measure:

$$M\sum_{k}^{i}K_i$$

where M is the measure and K is the set of normalized features. The set of normalized features are: 1+normalized journal impact factor, and normalized funding.

Table 18 Early detection results for measures weighted with normalized funding and average impact factor.

Results	ctf-iaf * fitness measure	citf-iaf * fitness-measure	ctf-sqrt-iaf*fitness measure	ctf*fitness measure
Bursts detected earlier than unmodified measure	1099(9%)	1(>1%)	590 (8%)	509(15%)

The results for the combined measure were almost identical to these results that just make use of funding. There is a slight improvement for the un-weighted measure, and CTF-IAF. However, there are fewer bursts identified by CTF-SQRT-IAF, than were found weighting CTF-SQRT-IAF with funding only.

CHAPTER 5 DISCUSSION

5.1 Burst Detection Framework Discussion

5.1.1 Characterizing Bursts

Burst characterization provides extremely important foundational knowledge for burst detection. For instance, approaches such as those of Tu et al. are based on specific assumptions about the life-cycle of bursts [81]. In doing so, they generated a model which did not account for the actually progression of emergent topics. For instance, their model assumed a single burst cycle, and only identified the first period at which a topic can be bursty, which is inappropriate in the scientific literature context as bursts can be cyclic.

In this work, an open discovery context was setup to evaluate bursts. Approaches such as those of Chen et al., use existing distinct datasets that include closely related documents and report the ability of the algorithm to find new scientific paradigms [55]. The advantage of the approach outlined in this dissertation is that it is not dependent on how thorough and up-to-date existing taxonomies are, allowing for a more complete topic model.

Section 4.2 demonstrates the differences in burst behavior by level of granularity. For instance, figures 13-18 illustrate that volatility decreases by level of frequency. Bursts last longer at lower levels of volatility. Bursts likely to be short give information about current/recent trends. Potentially long bursts indicate topics that can be planned around.

As show in table 9, weight is slightly negatively correlated with length. Kleinberg suggests that weight would be a good criterion for ranking. As discussed in section 4.2, weight may not be a good method for ranking bursts. My results indicate that the hierarchy could be used

to discover bursts through topic relatedness. For instance, documents in the overlapping hierarchy can be in multiple topics at multiple levels. Topic relatedness could be identified by grouping topics that share documents or that are sub-clusters of the same cluster. "Echocardiography for Assessment of Lv" is a topic that bursts for at least three years and has more than 30 documents associated with it during the planning horizon for experiments in section 4.3. It shares many documents with other clusters, which do not have topic bursts during that period, such as "Left Ventricular Ejection Fraction." A system could be built such that if a user did a search for the latter topic, they would be shown the former topic as a related bursty topic.

Figures 13-19 illustrates differences in bursty behavior by level of granularity. At the highest level, trend-lines are less volatile for the ten-year period. The highest level also has a limitation on the length of bursts discovered, with the maximum length of 4 years. This implies that if a burst at that level has been bursting for 3 years, it in very unlikely to burst for more than one more year. This information could be used at every level of granularity to find topics likely to burst in the future.

Reemergence information is very important. This is the reason why topics that have bursted in the past, are not removed from the list of possible bursty topics. An emerging topic detection application could alert uses of reemerging topics of interest.

5.1.2 Comparing Archival Weighting Measures

Archival weighting is a core component of my framework. My weightings performed better than the un-weighted Kleinberg method in several areas. These novel measures allow for a more comprehensive framework for burst detection.

Table 10 shows that CITF-IAF finds the highest number of bursts between 1997 and 2003 that are still bursting after 2003. One good potential criterion for a burst detection framework is that it helps identify bursts that stay bursty. If it could identify stable bursts, it could be used not only to understand the present, but also to pan for the future. This framework could, therefore, be used by researchers to plan a research agenda.

One significant motivator for this framework is that some topics that could be identified as bursty, are associated with well known important topics. The distribution of the number of clusters by IAF scores in figure 21 shows differences between the different measures. IAF scores indicate how common a term was in the historical period. The higher the IAF score the more common it was in the past. As illustrated in figure 21, CTF and DF-IAF finds the bursts with highest IAF. This means many of the bursty topics it finds are already quite common. Both measures identify "Cardiopulmonary Bypass" as a topic bursting during the period between 1997 and 2006, and it has the lowest IAF score of any topic identified as bursty. By this time, that was a well-known topic. CTF-SQRT-IAF, and CITF do not identify bursty topics as historically common as the aforementioned methods. That may make them more useful for finding bursty behavior that is not expected. CITF-IDF finds bursts at the lowest levels of historical frequency. These topics are ones that have never become highly frequent, and may be more interesting as a burst associated with these topics may be more surprising. In figure 21, the trend-lines are very similar for CTF-SQRT-IAF, CTF, and CTF-IAF, with the topic counts at the greater IAF ranges being higher in this order: CTF, CTF-SQRT-IAF, CTF-IAF. This implies that these three measures are behaving in a very similar way. The decision to use one over the other could be made based on how important IAF scores are.

5.1.4 Implications of Using a Topic Hierarchy

Generating a topic hierarchy allows for the automatic development of a vocabulary and has advantage over manual methods. For an emerging topic detection application this is very important. Emerging topics may not yet be added to MeSH or UMLS. Also, though those systems were not developed to be specific at the discipline level. Using a dynamically generated overlapping hierarchy allows for a more comprehensive and current topic structure at the discipline level.

Topic relatedness shows excellent potential as a tool for burst discovery. An application implementing the framework outlined in this work could support users entering search terms and finding bursty topics related to those terms. This would be a useful way to limit the number of bursty topics displayed to users, avoiding the pains of information overload.

A topic hierarchy would further allow policy makers to assess large-scale structural patterns and understand how certain initiatives fit within it. It would be very useful for initiative evaluation to see the level of each initiative's associated topics and whether they are bursty at the highest level. A topic hierarchy would be useful for researchers and clinicians as well. Researchers are often faced with solving complex problems, and perform many interrelated queries of literature search engines. Understanding topic bursts in relation to other topics would be very useful. A research may want to know if a given topic burst has bursty related topics that are emerging at a higher level of intensity. A researcher may decide to investigate those topics *because* they are bursty to determine if that increased burstiness correlates to more promising methodology, conceptual develop or techniques.

5.1.4 Measuring Performance at Different Levels of Granularity

Identifying typical burst behaviors at various levels of granularity can help policy makers, clinicians, and researchers assess where a given topic is in its life cycle. If policy makers want to fund emergent research that is somewhat stable, burst characterization data would help. Decision-making about whether to fund portfolios associated with unstable but potentially impactful bursty topics could also be supported. If researchers are trying to reference work outside their area of expertise they may want to stick to work that is current and associated with a relatively established area.

Citation counts for clusters were used to determine the extent to which the topics were impactful in the future. As show in table 11, at the highest level, CTF-IAF gets the highest average citation count. For every other level of granularity, CITF_IAF has the highest average citation count. This indicates that these measures do not just find bursty topics, but topics with significant impact when compared to the un-modifed Kleinberg method.

One issue with each of the different measures is the amount of noise reduction that must be performed to get a similarly short list of candidate bursts for review. Both CTF-SQRT-IAF and CTF-IAF identify more bursty topics than CTF. CTF-IAF and CTF identify the same number of bursts at the highest level of granularity. After noise reduction was performed, bursts identified by the criteria described in section 4.3 still included many highly frequent bursts for CTF-IAF – considerably more than for any other measure. However, with that criterion, CITF-IAF identified the most manageable number of bursts, 143. Each of the other three measures identified more than 1000 bursts, which will require further noise reduction in a real-time use case.

5.1.5 Early Detection with the Fitness Model

The fitness score was developed to help identify potentially more important bursts by indicating a burst's likeliness to continue in frequency. Unfortunately, normalized h-index could be used for the combined measure. As shown in table 12, this score was able to identify 22% to 41% of bursts early for each weighted measure with the exception of CITF-IAF. The score, described in section 4.4.3, using funding data does much better than when using impact factor. Adding the two factors together only gave a slight improvement for CTF. It is interesting that normalized impact factor has little effect. Hence, it is not just topics associated with top journals that burst. Funding has a great impact on what is studied, as researchers are dependent on funding opportunities to do research. This information can be used not only to find important bursts, but also to understand the development of research topics.

To find bursts one can use the effect of the fitness measure to find prominent bursts. The difference in burst strength and length after using this score can be use to find the most promising bursts. The fitness score that improves burst detection the most provides information regarding how bursts develop. If funding has a greater impact than impact factor it indicates that it is a far greater driver for research development.

5.1.6 Evaluation Methods

As noted earlier, developing an appropriate evaluation for emerging detection algorithms has proved difficult. Coming up with a set of criteria for a burst detection algorithm is not easy. Several other researchers use experts to determine how well their algorithms perform. However, it is difficult for even experts to stay abreast of trends, so their input is of limited value. Some of the approaches do not rely on characterizing the burstiness of previously identified emerging topic

drift. Mane et al.'s method analyzing the PNAS dataset is one [5]. However, the authors were only interested in most frequent terms. That may not be useful as those are the most large scale-trends, and don't provide users with new information. If a trend is at a large enough scale, one may safely assume most researchers within the discipline know about it. The new and emergent trends would be the most significant.

As mentioned in chapter 1, there are differences between a topic bursting in science as opposed to news. In science, a topic remaining viable at a consistent level of burstiness can still be associated with novelty and interest in the community. Therefore, while comparing the slope of bursty topics is important, it may not be as important as an evaluative tool as others. CTF-SQRT-IAF and CTF-IAF both had smaller average slopes for each level of granularity than CTF. However, if a topic experiences a burst but does not increase that may still be novel, it may just be that there are not many researchers dedicating themselves to the topic or that it takes a long time for developments to take place. For that reason the difference in slope between these measure does not mean they should be ruled out as relevant. In fact, CTF-SQRT-IAF has the highest average r-squared, indicating that it is identify trends that can be better modeled. This may indicate that it identifies bursts that behave in a manner more easily predicatable. CITF had the highest average slope for the lower two levels of granularity. This indicates that it does a good job at that level of finding bursts with an upward trend.

Method comparison based on future impact, linear trend, and burst length provides information on bursts that are useful. Researchers looking to find bursty topics would benefit from finding bursts that have a significant impact, as it increases the likelihood that their work will have a lasting impact. Topics with a linear trend would indicate that the burst might get to a higher level of intensity. Finding topics of greater burst length indicates that a burst may be

stable for several years. Finding topics with the potential to burst for a few more years would be very useful for those conducting research.

5.2 Parameterizing the Burst Detect Framework for Various Use Cases

Section 2.1 and 2.2 detailed many of the needs of policy makers, researchers and clinicians. The emerging topic detection algorithms reviewed in section 2 are insufficient in deal with them. Both clinicians and researchers need to find not just emerging research, but research relevant to them. Policy makers need tools to help them determine which portfolios to fund, and need to know whether a topic is emergent.

As mentioned in section 2.1, policy makers fund high-risk high-rewards emerging research projects. Knowing which bursts may be more likely to stay emergent would also be helpful as policy makers do not want to fund too many projects which don't pan out. For that reason this framework includes a fitness model to assess whether bursts are likely to stay bursty. Researchers, benefit as they can associate themselves with an emerging area, leading to more visibility, greater impact, and publication opportunities. Clinicians could benefit, as they could stay up-to-date on current trends.

5.2.1 Implications for Policy

My framework has clear potential to assist policy makers based on their information needs, and provides interpretive flexibility. An emerging topic detection application would not require the understanding of bibliometric methods and would use the publicly available MEDLINE data set, making it easy to use.

In section 2.1, two issues funding agencies like NIH and NSF have were identified. The first issue is the identification of promising high-risk high rewards research. The second issue is the evaluation of initiatives aimed at increasing the speed of knowledge production.

The first issue can be addressed through the application of IDF-IAF scoring. High-risk high-rewards research is transformative research with the potential for high impact. As mentioned in section 2.1, there are difficulties selecting promising high-risk high-rewards research. Identifying potentially transformative research is difficult. It is difficult because it can only truly be identified in retrospect, and many of the projects that get funded have little or no long term utility. Using IDF-IAF could help improve the search for potentially transformative research. This score finds bursts of longest length, for topics that are not only infrequent in the past, but also infrequent in the present. Therefore it can identify topics that are associated with topics that have not developed enough to reach the mainstream, and are potentially high-risk. Giving precedent to topics associated with IDF-IAF can be beneficial for several reasons. First, a topic bursting, even on a small scale, on a small scale is likely an indication that a promising discovery has been made and the research community is responding to it. As bursty topics identified with IDF-IAF tend to burst for a long period, it can indicate that research projects associated with those topics *would* be successful as they are likely to both continue in frequency and be associated with a high average citation count. The topic hierarchy could be used to identify bursty topics of potential impact. The topic hierarchy provides relationship information between topics. A rare bursty topic related to a more prominent research area, such as cancer, for which a new treatment could have a huge impact, scould be identified with the topic hierarchy. Using this measure may reduce the risk of funding high -risk high-rewards research that does not produce results.

The second issue identified in section 2.1 is the problem of initiative evaluation. NIH has recently focused on awarding large grants to large research teams using cutting-edge technology to speed up the process of novel medical research output. A burst in a topic is likely an indication of an important discovery which has propelled a given research area. NIH could use my framework to evaluate research by assessing the extent to which initiatives funded are associated with bursty, and potentially novel medical research. Consider the CTSA awards, which are among the biggest grants given to universities. If the 60-some institutions receiving these awards are associated with large-scale bursty topics, there may be cause to view the program as successful. Temporal analysis could be used to determine if there was an increase in association with bursty topics over time. Using CTF-SQRT-IAF would help reduce the number of historically frequent topics identified, (it do not find bursty topics with quite as high of IAF scores as CTF, or CTF-IAF). Compared to CITF-IAF, it finds topics that are not necessarily so rare, as to limit its potential at different levels of granularity. By generating levels of granularity for bursts, a richer exploration of bursty topics can be performed. Large-scale multi-university initiatives may be associated with large-scale bursts or many small-scale bursts. It may be that the desired effect is to create large-scale burst in a specific area. However, without knowing how to quantify the range for large-scale bursts, that will not be possible. Again, the most frequent topics do not burst. Understanding the behavior of bursts at different levels of granularity, and the frequency of bursts at each level is important. Also, individual research sites may be associated with small-scale bursts, which may not seem very impressive. However, knowing that most bursts are small scale, and if the fitness score could be used to suggest that some of these bursts might be more significant in the future, that will assist evaluators. Also, knowing what is

typical for research groups at each level of granularity can help evaluators know who is reaching the base-line and who is exceeding expectations.

Policy makers typically use peer review to for the tasks mentioned. There are many potential benefits to using bibliometric methods instead. Peer-review suffers from subjectivity, and high cost. In some cases, replacing peer-review with bibliometric methods for policy decision-making has proven efficient at a guaranteed lower cost. Though bibliometrics have great potentially in this area, they are not frequently used. This is in part do to the fact that many bibliometric methods require a specialized skill-set, or the necessary applications like to data may not be available. Many bibliometric methods lack interpretive flexibility. An application could be built using this framework that would not require a specialized skill-set, only an understanding of the type of emerging topic of interest (e.g. historically rare and currently rare). It would be dynamic enough to answer many different types of questions.

5.2.3 Improving Tools for Clinicians and Medical Researchers

Researchers have different needs than policy makers and the framework could be parameratized in a different way to meet their needs. Biomedical researchers as tasked wit solving complex problems that require planning, and an understanding of the relationship between concepts. Researchers of often confronted with what is referred to as *weak problem solving,* which is associated with a vague understanding of the problem space, and an inability to come up with a systematic plan to resolve information needs. The issues can be exacerbated with the difficulty of determining what is most current in a given research area. A researchers carrer prospect can be effected by their success at these takes. The fitness score in conjunction with CTF-SQRT-IAF would be very useful in assisting with these tasks, as it identifies bursts

with trend-lines easiest to model and it does not identify bursts as historically common as CTF or CTF-IAF. CTF-SQRT-IAF identifies more topics than CITF. Researchers typically work on topics where they have expertise. Limiting the number of bursts found by too large a degree may decrease the frameworks usefulness, as identifying too few bursty topics may mean most researchers cannot find bursty topics related to their area. Researchers face difficulty determining which topics are currently emerging. At various levels of granularity, researchers may want to find bursty topics. A topic bursting at the lowest level of granularity, can give a researcher an opportunity to make a huge impact in a slowly growing research area. Alternatively, if the research area of interest is associated with a burst at a more coarse grain topic level, they may be able to publish more as that area may have several interrelated topics that are bursting, and creating the burst at the more coarse grain topic level. Doing work associated with that bursty topic at that level of granularity might give them the opportunity to have more projects, which can have an impact. Using the fitness score would be very beneficial for researchers who want to publish on a topic before it becomes too popular and well known, allowing them to make an impact when the area is still forming. Topics identified as bursting based on the fitness score can help indicate future bursty behavior. Topics whose length or burst strength is increased by use of the fitness score would make very promising candidate topics for analysis.

An emerging topic detection application, generating information on new developments by discipline, specific to a given research area, would be instrumental in the development of a point of care toolkit. The framework presented in this work could be used to enhance up-to-date tools for use by clinicians. As mentioned in section 2.2, clinicians have very limited time during the course of the day to answer clinical questions and rely on summaries and practice guidelines, regardless of whether those resources are evidence based. In the scientific literature, there have

been calls information tools, which alert clinicians to new, relevant, and valid information; tools which tailor information to the appropriate specialty of each physician. A topic's history could be used to alert clinicians to new discoveries and allow them to search for new treatments. If a topic burst for a significant amount of time in the recent past, was associated with large levels of funding, and had high citation counts it would indicate that it was considered an important area, and potentially, that the burstyness is a sign of novel research. Information provided could be tailored to the specialty of each physician by using the topic hierarchy to find bursty topics related to specialty areas. Using the framework to find such work would allow for better identification of emergent topics that could help clinicians make decisions.

CHAPTER 6 CONCLUSIONS AND FUTURE WORK

This dissertation offers a novel framework for emerging topic detection in biomedicine. My method is more robust than other methods for several reasons, resulting in a framework that takes the typical life-cycle of bursts into consideration. Instead of making assumptions about bursts, they were characterized from a representative corpus drawn from the biomedical literature. Characterizing bursts supported the identification of differences between news topic burst life-cycles, and scientific topic burst life-cycle. It also allowed for the identification of differences between bursts at different levels of granularity. These differences helped shape this framework so that it was more comprehensive than other methods.

There are many aspects of my model that are novel, and produce good results. First, the topic hierarchy was generated using clustering techniques. Modeling the topic hierarchy in this way has many advantages over using curated taxonomies that other researchers in this area have used. Using cluster analysis allows for the identification of terms that may not be included in annotated taxonomies, as these terms are new or not considered as relevant at the time the taxonomy was last updated. Second, weighting topics by historical frequency allows for better identification of bursts that are associated with less well-known areas, and therefore more surprising. If identified bursts are associated with stable topics, the framework may be informing users of what they already know. Third, the fitness score allows for the early identification of bursty terms. This feature can help researchers stay on the cutting edge. These three aspects of the framework presented in this work make it a robust, novel, and useful framework for the identification of bursty terms.

Using this framework as a basis, there are numerous research projects that would further the understanding of the structure of scientific development. Not only could these research projects be used to understand this structure, but also to improve the framework outlined.

This framework could be used to compare different disciplines. Burst characterization may be different for different fields, and different areas could benefit from a different parameterization. This would allow for a more robust, comprehensive framework, and would make the framework more beneficial to a larger group. Researchers and clinicians could find results that would more specifically help them with their particular research area, and policy makers could use differences between disciplines to inform decision-making.

This framework could be improved by generating a more dynamic historical period. For the research described in this dissertation, the historical period used to develop archival scores was static. It was developed so that it ended before the first year in the planning horizon. A sliding window for the historical period would further contextualize the frequency counts for each year.

Another aspect of the framework that could benefit from dynamic modeling is the parameterization of the fitness score. For instance, funding patterns change over time. At different times, weighting frequency counts with associated funding information can have differing effects on early burst detection. Dynamically parameterizing the fitness score could be achieved by identifying how well each feature performs in 3-5 year sliding windows. This could improve the results of my framework.

My framework could be used in conjunction with algorithms other thank Kleinberg's. The algorithm used is really just one feature of the framework. The topic hierarchy, historical

weightings, and fitness score could all be used with other emerging detection algorithms, if another one was deemed more useful for a giving use-case. The framework is flexible enough to be used independently of any particular algorithm.

This dissertation analyzed the single discipline of cardiology. This framework could be used to identify bursts on the combined topic hierarchies from different disciplines. This could be done for all clinical disciplines, just surgery-related disciplines, or only basic scientists, for example. This would allow for further stratification of research topics at various levels of granularity. Large-scale structural analysis could be used not only for burst detection, but also for planning a research agenda at an institutional or national level.

An obvious research project to pursue using this framework involves achieving a greater understanding of translational medicine. As mentioned in section 2.1, NIH has focused considerable efforts to support and speed up the process of translation of basic to clinical research. This framework could be used to further elucidate translational research trends. An analysis of the projection of basic science bursty topics into clinical research could achieve this. If the literature was initially stratified into basic and clinical science this approach could detect translational even in the absence of citation linking the two. One could analyze clinical science research to determine the typical time period for translation of bursty topics from basic to clinical. This framework could also be used to identify current translation trends by discipline. This would inform policy makers trying to understand current trends to determine what work needs to be done.

Finally, the effectiveness of this framework could be further evaluated with the use of user reviews. Evaluations from policy makers, medical researchers, and clinicians could help

determine which aspects of the framework are most useful, and what could be done to improve the model.

REFERENCES

[1] T. S. Kuhn and D. Hawkins, “The Structure of Scientific Revolutions,” Am. J. Phys., vol. 31, no. 7, pp. 554–555, Jul. 1963.

[2] A. F. J. van Raan, “On Growth, Ageing, and Fractal Differentiation of Science,” Scientometrics, vol. 47, no. 2, pp. 347–362, Feb. 2000.

[3] A. H. Renear and C. L. Palmer, “Strategic Reading, Ontologies, and the Future of Scientific Publishing,” Science, vol. 325, no. 5942, pp. 828–832, Aug. 2009.

[4] D. He, X. Zhu, and D. S. Parker, “How Does Research Evolve? Pattern Mining for Research Meme Cycles,” in Data Mining, IEEE International Conference on, Los Alamitos, CA, USA, 2011, vol. 0, pp. 1068–1073.

[5] K. K. Mane and K. Börner, “Mapping topics and topic bursts in PNAS,” Proc. Natl. Acad. Sci., vol. 101, no. suppl 1, pp. 5287–5290, Apr. 2004.

[6] J. Kleinberg, “Bursty and hierarchical structure in streams,” Data Min. Knowl. Discov., vol. 7, no. 4, pp. 373–397, 2003.

[7] Y.-H. Tseng, Y.-I. Lin, Y.-Y. Lee, W.-C. Hung, and C.-H. Lee, “A comparison of methods for detecting hot topics,” Scientometrics, vol. 81, no. 1, pp. 73–90, Mar. 2009.

[8] F. S. Collins, “Reengineering Translational Science: The Time Is Right,” Sci. Transl. Med., vol. 3, no. 90, pp. 90cm17–90cm17, Jul. 2011.

[9] Sung NS, Crowley, Jr WF, Genel M, and et al, “Central challenges facing the national clinical research enterprise,” JAMA, vol. 289, no. 10, pp. 1278–1287, Mar. 2003.

[10] R. M. Califf and L. Berglund, “Linking Scientific Discovery and Better Health for the Nation: The First Three Years of the NIH’s Clinical and Translational Science Awards:,” Acad. Med., vol. 85, no. 3, pp. 457–462, Mar. 2010.

[11] A. Elzinga, “Features of the current science policy regime: Viewed in historical perspective,” Sci. Public Policy, vol. 39, no. 4, pp. 416–428, Aug. 2012.

[12] G. Abramo, C. A. D'Angelo, and A. Caprasecca, "Allocative efficiency in public research funding: Can bibliometrics help?," Res. Policy, vol. 38, no. 1, pp. 206–215, 2009.

[13] D. Hicks, H. Tomizawa, Y. Saitoh, and S. Kobayashi, "Bibliometric techniques in the evaluation of federally funded research in the United States," Res. Eval., vol. 13, no. 2, pp. 76–86, Aug. 2004.

[14] J. Lane, "Let's make science metrics more scientific," Nature, vol. 464, no. 7288, pp. 488–489, Mar. 2010.

[15] I. Rafols, A. L. Porter, and L. Leydesdorff, "Science overlay maps: A new tool for research policy and library management," J. Am. Soc. Inf. Sci. Technol., vol. 61, no. 9, pp. 1871–1887, Sep. 2010.

[16] C. S. Wagner and J. Alexander, "Evaluating transformative research programmes: A case study of the NSF Small Grants for Exploratory Research programme," Res. Eval., vol. 22, no. 3, pp. 187–197, Sep. 2013.

[17] H. Grupp and M. E. Mogee, "Indicators for national science and technology policy: how robust are composite indicators?," Res. Policy, vol. 33, no. 9, pp. 1373–1384, Nov. 2004.

[18] S. Scotchmer, Innovation and incentives. MIT press, 2004.

[19] E. A. Zerhouni, "Clinical research at a crossroads: the NIH roadmap," J. Investig. Med., vol. 54, no. 4, pp. 171–173, 2006.

[20] E. A. Zerhouni and B. Alving, "Clinical and Translational Science Awards: a framework for a national research agenda," Transl. Res., vol. 148, no. 1, pp. 4–5, 2006.

[21] S. Wuchty, B. F. Jones, and B. Uzzi, "The increasing dominance of teams in production of knowledge," Science, vol. 316, no. 5827, pp. 1036–1039, 2007.

[22] J. Esparza and T. Yamada, "The discovery value of 'Big Science,'" J. Exp. Med., vol. 204, no. 4, pp. 701–704, 2007.

[23] D. Stokols, K. L. Hall, B. K. Taylor, and R. P. Moser, "The Science of Team Science: Overview of the Field and Introduction to the Supplement," Am. J. Prev. Med., vol. 35, no. 2, Supplement 1, pp. S77–S89, Aug. 2008.

[24] M. Wehling, “Translational medicine: science or wishful thinking?,” J. Transl. Med., vol. 6, no. 1, p. 31, 2008.

[25] G. Ordonez-Matamoros, S. E. Cozzens, and M. Garcia-Luque, “International co-authorship and research team performance in Colombia,” in 2009 Atlanta Conference on Science and Innovation Policy, 2009, pp. 1–9.

[26] B. L. Ponomariov and P. C. Boardman, “Influencing scientists’ collaboration and productivity patterns through new institutions: University research centers and scientific and technical human capital,” Res. Policy, vol. 39, no. 5, pp. 613–624, Jun. 2010.

[27] C. L. Palmer, M. H. Cragin, and T. P. Hogan, “Weak information work in scientific discovery,” Inf. Process. Manag., vol. 43, no. 3, pp. 808–820, 2007.

[28] C. A. Bana e Costa and M. D. Oliveira, “A multicriteria decision analysis model for faculty evaluation,” Omega, vol. 40, no. 4, pp. 424–436, Aug. 2012.

[29] M. Fiszman, D. Demner-Fushman, H. Kilicoglu, and T. C. Rindflesch, “Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation,” J. Biomed. Inform., vol. 42, no. 5, pp. 801–813, 2009.

[30] W. Xuan, M. Dai, B. Mirel, J. Song, B. Athey, S. Watson, and F. Meng, “Open Biomedical Ontology-based Medline exploration,” BMC Bioinformatics, vol. 10, no. Suppl 5, p. S6, 2009.

[31] Y. Lin, W. Li, K. Chen, and Y. Liu, “A document clustering and ranking system for exploring MEDLINE citations,” J. Am. Med. Inform. Assoc., vol. 14, no. 5, pp. 651–661, 2007.

[32] H. T. Zheng, C. Borchert, and H. G. Kim, “GOClonto: An ontological clustering approach for conceptualizing PubMed abstracts,” J. Biomed. Inform., vol. 43, no. 1, pp. 31–40, 2010.

[33] P. Zweigenbaum, “Knowledge and reasoning for medical question-answering,” in Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions, 2009, pp. 1–2.

[34] S. R. Jonnalagadda, G. Del Fiol, R. Medlin, C. Weir, M. Fiszman, J. Mostafa, and H. Liu, “Automatically extracting sentences from Medline citations to support clinicians’ information needs,” J. Am. Med. Inform. Assoc., vol. 20, no. 5, pp. 995–1000, 2013.

[35] A. Névéol and Z. Lu, “Automatic integration of drug indications from multiple health resources,” in Proceedings of the 1st ACM International Health Informatics Symposium, 2010, pp. 666–673.

[36] L. M. Schilling, J. F. Steiner, K. Lundahl, and R. J. Anderson, “Residents’ patient-specific clinical questions: opportunities for evidence-based learning,” Acad. Med., vol. 80, no. 1, pp. 51–56, 2005.

[37] M. L. Green, M. A. Ciampi, and P. J. Ellis, “Residents’ medical information needs in clinic: are they being met?,” Am. J. Med., vol. 109, no. 3, pp. 218–223, 2000.

[38] F. A. I. Riordan, E. M. Boyle, and B. Phillips, “Best paediatric evidence; is it accessible and used on-call?,” Arch. Dis. Child., vol. 89, no. 5, pp. 469–471, 2004.

[39] W. Putnam, P. L. Twohig, F. I. Burge, L. A. Jackson, and J. L. Cox, “A qualitative study of evidence in primary care: what the practitioners are saying,” Can. Med. Assoc. J., vol. 166, no. 12, pp. 1525–1530, 2002.

[40] D. C. Slawson, “Teaching evidence-based medicine: should we be teaching information management instead?,” Acad. Med., vol. 80, no. 7, p. 685, 2005.

[41] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, “Topic Detection and Tracking Pilot Study Final Report,” Comput. Sci. Dep., Feb. 1998.

[42] D. Eichmann, M. Ruiz, P. Srinivasan, N. Street, C. Culy, and F. Menczer, “A cluster-based approach to tracking, detection and segmentation of broadcast news,” in Proceedings of the DARPA Broadcast News Workshop, 1999, pp. 69–76.

[43] C. Clifton, R. Cooley, and J. Rennie, “TopCat: data mining for topic identification in a text corpus,” IEEE Trans. Knowl. Data Eng., vol. 16, no. 8, pp. 949–964, 2004.

[44] A. Pons-Porrata, R. Berlanga-Llavori, and J. Ruiz-Shulcloper, “Topic discovery based on text mining techniques,” Inf. Process. Manag., vol. 43, no. 3, pp. 752–768, May 2007.

[45] M. Khalilian and N. Mustapha, “Data Stream Clustering: Challenges and Issues,” arXiv:1006.5261, Jun. 2010.

[46] J. Yao, B. Cui, Y. Huang, and Y. Zhou, “Bursty event detection from collaborative tags,” World Wide Web, vol. 15, no. 2, pp. 171–195, 2012.

[47] I. Suba\vsic and B. Berendt, “From bursty patterns to bursty facts: The effectiveness of temporal text mining for news,” in Proceeding of the 2010 conference on ECAI, 2010, pp. 517–522.

[48] J. Ratkiewicz, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani, “Traffic in social media ii: Modeling bursty popularity,” in Social Computing (SocialCom), 2010 IEEE Second International Conference on, 2010, pp. 393–400.

[49] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, “On the bursty evolution of blogspace,” World Wide Web, vol. 8, no. 2, pp. 159–178, 2005.

[50] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim, “Finding bursty topics from microblogs,” in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, 2012, pp. 536–544.

[51] S. L. Decker, “Detection of bursty and emerging trends towards identification of researchers at the early stage of trends,” Aug-2007. [Online]. Available: http://athenaeum.libs.uga.edu/handle/10724/9958. [Accessed: 01-Mar-2012].

[52] W. Ke, K. Borner, and L. Viswanath, “Major Information Visualization Authors, Papers and Topics in the ACM Library,” in IEEE Symposium on Information Visualization, 2004. INFOVIS 2004, 2004, pp. r1–r1.

[53] K. W. Boyack, K. Mane, and K. Borner, “Mapping Medline papers, genes, and proteins related to melanoma research,” in Eighth International Conference on Information Visualisation, 2004. IV 2004. Proceedings, 2004, pp. 965–971.

[54] S. Morinaga and K. Yamanishi, “Tracking Dynamics of Topic Trends Using a Finite Mixture Model,” New York, NY, USA, 2004, pp. 811–816.

[55] C. Chen, “Measuring the movement of a research paradigm,” in Proc. SPIE 5669, 2005, vol. 63.

[56] Y. Chi, B. L. Tseng, and J. Tatemura, “Eigen-trend: trend analysis in the blogosphere based on singular value decompositions,” in Proceedings of the 15th ACM international conference on Information and knowledge management, New York, NY, USA, 2006, pp. 68–77.

[57] D. He and D. S. Parker, “Topic dynamics: an alternative model of bursts in streams of topics,” in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 443–452.

[58] V. Larivière and Y. Gingras, “The impact factor’s Matthew Effect: A natural experiment in bibliometrics,” J. Am. Soc. Inf. Sci. Technol., vol. 61, no. 2, pp. 424–427, 2010.

[59] R. S. J. Tol, “The Matthew effect defined and tested for the 100 most prolific economists,” J. Am. Soc. Inf. Sci. Technol., vol. 60, no. 2, pp. 420–426, 2009.

[60] J. J. Murphy, Technical analysis of the financial markets: A comprehensive guide to trading methods and applications. Penguin, 1999.

[61] F. Moerchen, D. Fradkin, M. Dejori, and B. Wachmann, “Emerging trend prediction in biomedical literature,” in AMIA Annual Symposium Proceedings, 2008, vol. 2008, p. 485.

[62] S. A. Morris, G. Yen, Z. Wu, and B. Asnake, “Time line visualization of research fronts,” J. Am. Soc. Inf. Sci. Technol., vol. 54, no. 5, pp. 413–422, 2003.

[63] Y. Takeda and Y. Kajikawa, “Optics: A bibliometric approach to detect emerging research domains and intellectual bases,” Scientometrics, vol. 78, no. 3, pp. 543–558, 2009.

[64] N. Shibata, Y. Kajikawa, Y. Takeda, and K. Matsushima, “Detecting emerging research fronts based on topological measures in citation networks of scientific publications,” Technovation, vol. 28, no. 11, pp. 758–775, Nov. 2008.

[65] S. A. Morris, “Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution,” J. Am. Soc. Inf. Sci. Technol., vol. 56, no. 12, pp. 1250–1273, 2005.

[66] H. Small and P. Upham, “Citation structure of an emerging research area on the verge of application,” Scientometrics, vol. 79, no. 2, pp. 365–375, May 2009.

[67] F. \AAström and P. Sweden, “Changes in the LIS research front: Time-sliced co-citation analyses of LIS journal articles.”

[68] H. G. Small, “Cited Documents as Concept Symbols,” Soc. Stud. Sci., vol. 8, no. 3, pp. 327–340, Aug. 1978.

[69] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering: an overview,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 2, no. 1, pp. 86–97, Jan. 2012.

[70] P. Yu and H. Van de Sompel, “Networks of scientific papers,” Science, vol. 169, pp. 510–515, 1965.

[71] E. GARFIELD, “From Citation Indexes to Informetrics: Is the Tail Now Wagging the Dog ?,” Libri, vol. 48, no. 2, pp. 67–80, 2009.

[72] M. J. Tobin, “Remembrance of weaning past: the seminal papers,” Intensive Care Med., vol. 32, no. 10, pp. 1485–1493, Oct. 2006.

[73] P. Ahlgren and C. Colliander, “Document-document similarity approaches and science mapping: Experimental comparison of five approaches,” J. Informetr., vol. 3, no. 1, pp. 49–63, Jan. 2009.

[74] D. Knights, M. C. Mozer, and N. Nicolov, “Detecting Topic Drift with Compound Topic Models.,” in ICWSM, 2009.

[75] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles, “Detecting topic evolution in scientific literature: how can citations help?,” in Proceedings of the 18th ACM conference on Information and knowledge management, New York, NY, USA, 2009, pp. 957–966.

[76] T. Qian, P. C. Y. Sheu, S. Li, and L. Wang, “A Scientific Theme Emergence Detection Approach Based on Citation Graph Analysis,” in Tools with Artificial Intelligence, 2008. ICTAI’08. 20th IEEE International Conference on, 2008, vol. 2, pp. 269–273.

[77] R. Schult and M. Spiliopoulou, “Discovering emerging topics in unlabelled text collections,” in Advances in Databases and Information Systems, 2006, pp. 353–366.

[78] Q. Mei and C. Zhai, “Discovering evolutionary theme patterns from text: an exploration of temporal text mining,” in Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, New York, NY, USA, 2005, pp. 198–207.

[79] W. Mendenhall and T. Sincich, “A second course in statistics,” 1996.

[80] C. Buckely, “Text REtrieval Conference (TREC) trec_eval IR evaluation package.” [Online]. Available: http://trec.nist.gov/trec_eval/. [Accessed: 23-Oct-2013].

[81] Y. N. Tu and J. L. Seng, “Indices of novelty for emerging topic detection,” Inf. Process. Manag., vol. 48, no. 2, pp. 303–325, 2012.

[82] C. C. Chen, Y. T. Chen, and M. C. Chen, “An aging theory for event life-cycle modeling,” Syst. Man Cybern. Part Syst. Hum. IEEE Trans. On, vol. 37, no. 2, pp. 237–248, 2007.

[83] Y. Jo, C. Lagoze, and C. L. Giles, “Detecting research topics via the correlation between graphs and texts,” in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, 2007, pp. 370–379.

[84] P. Yin and R. Cui, “Evaluation of literature frontier based on latent semantic analysis,” in 2012 IEEE Symposium on Robotics and Applications (ISRA), 2012, pp. 403–406.

[85] D. D. S. Price, “A general theory of bibliometric and other cumulative advantage processes,” J. Am. Soc. Inf. Sci., vol. 27, no. 5, pp. 292–306, 1976.

[86] W. Ke, “A fitness model for scholarly impact analysis,” Scientometrics, vol. 94, no. 3, pp. 981–998, Mar. 2013.

[87] A.-L. Barabási, “Scale-free networks: a decade and beyond,” Science, vol. 325, no. 5939, pp. 412–413, 2009.

[88] D. Wang, C. Song, and A.-L. Barabási, “Quantifying Long-Term Scientific Impact,” Science, vol. 342, no. 6154, pp. 127–132, Oct. 2013.

[89] R. Costas and M. Bordons, “The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level,” J. Informetr., vol. 1, no. 3, pp. 193–203, Jul. 2007.

[90] S. Osiriski, J. Stefanowski, and D. Weiss, “Lingo: Search results clustering algorithm based on singular value decomposition,” in Intelligent information processing and web mining: proceedings of the International IIS: IIPWM’04 Conference held in Zakopane, Poland, May 17-20, 2004, 2004, p. 359.

[91] O. Bodenreider, “The Unified Medical Language System (UMLS): integrating biomedical terminology,” Nucleic Acids Res., vol. 32, no. Database issue, pp. D267–70, Jan. 2004.

[92] B. L. Humphreys and D. A. Lindberg, “The UMLS project: making the conceptual connection between users and the information they need.,” Bull. Med. Libr. Assoc., vol. 81, no. 2, p. 170, 1993.

[93] J. Janruang and W. Kreesuradej, “A new web search result clustering based on true common phrase label discovery,” in Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on, 2007, p. 242.

[94] S. Osiński and D. Weiss, “Conceptual Clustering Using Lingo Algorithm: Evaluation on Open Directory Project Data, Advanced in Soft Computing, Intelligent Information Processing and Web Mining,” in Proceedings of the the International IIS: IIPWM’04 Conference, pp. 369–378.

[95] S. Osiński and D. Weiss, “Carrot2: Design of a Flexible and Efficient Web Information Retrieval Framework,” in Advances in Web Intelligence, P. S. Szczepaniak, J. Kacprzyk, and A. Niewiadomski, Eds. Springer Berlin Heidelberg, 2005, pp. 439–444.

[96] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, “A comparison of extrinsic clustering evaluation metrics based on formal constraints,” Inf. Retr., vol. 12, no. 4, pp. 461–486, Jul. 2008.

[97] K. Premalatha and A. M. Natarajan, “A literature review on document clustering,” Inf. Technol. J., vol. 9, no. 5, pp. 993–1002, 2010.

[98] C. Blattman, J. Hwang, and J. G. Williamson, “Winners and losers in the commodity lottery: The impact of terms of trade growth and volatility in the Periphery 1870–1939,” J. Dev. Econ., vol. 82, no. 1, pp. 156–179, Jan. 2007.

[99] J. E. Hirsch, “An index to quantify an individual’s scientific research output,” Proc. Natl. Acad. Sci. U. S. A., vol. 102, no. 46, p. 16569, 2005.

[100] W. Glänzel, “On the h-index-A mathematical approach to a new measure of publication activity and citation impact,” Scientometrics, vol. 67, no. 2, pp. 315–321, 2006.

APPENDIX

Words added to the stop-word list for the STC and Lingo Clustering Algorithms

study	improve	enhanced
studies	previously	impaired
decreased	conclusion	impair
decrease	additional	combined
increase	presented	combine
increased	purpose	overall
hypothesis	defined	alter
investigate	tested	altered
statistically	better	experience
useful	remains	reduce
analyze	conclude	history
analyzed	remained	contribute
indicate	review	given
association	second	result
Comparison	described	resulted
Considered	degree	Increasing
techniques	elevated	direct
seen	efficacy	produce
Comparison	randomized	produced
Considered	show	property
techniques	calculated	properties
seen	various	rapid
characteristics	died	course

onset	healthy	maximal
subsequent	absence	successfully
diagnostic	specific	good
implemented	despite	examination
influence	compare	surface
population	compared	0.5
implanted	clinical	year
procedures	associated	weight
procedure	results	median
maximum	mean	intervention
involve	treatment	resulting
involved	effects	recently
extent	blood	prior
approximately	effect	groups
pattern	right	analysis
Regression	used	data
Complex	performed	time
Significant	control	age
Significantly	detected	patient
correlated	performance	years

Patterns added to the stop-label list for the STC and Lingo Clustering Algorithms

(?i)0.*	(?i)8.*	(?i)\d+
(?i)1.*	(?i)9.*	(?i)acute
(?i)28.*	(?i)recently	(?i).*\d+.*
(?i)3.*	(?i)limited	(?i).*group.*
(?i)4.*	(?i)median	(?i).*groups.*
(?i)5.*	(?i)year	(?i).*patient.*
(?i)6.*	(?i)Resulting	
(?i)7.*	(?i)\d+.*	