This is a preprint of an article published in **Chaomei Chen (2011) Predictive Effects of Structural Variation on Citation Counts. *Journal of the American Society for Information Science*. doi: 10.1002/asi.21694**. copyright © 2011 (American Society for Information Science and Technology). This preprint has been updated to reflect changes in the final version. http://dx.doi.org/10.1002/asi.21694

Predictive Effects of Structural Variation on Citation Counts

Chaomei Chen

College of Information Science and Technology

Drexel University
3141 Chestnut Street
Philadelphia, PA 19104 USA
+1 215 895 6627
chaomei.chen@drexel.edu

ABSTRACT

A critical part of a scientific activity is to discern how a new idea is related to what we know and what may become possible. As the number of new scientific publications arrives at a rate that rapidly outpaces our capacity of reading, analyzing, and synthesizing scientific knowledge, we need to augment ourselves with information that can guide us through the rapidly growing intellectual space effectively. In this article, we address a fundamental issue concerning with what information may serve as early signs of potentially valuable ideas. In particular, we are interested in information that is routinely available and derivable upon the publication of a scientific paper without assuming the availability of additional information such as its usage and citations. We propose a theoretical and computational model that predicts the potential of a scientific publication in terms of the degree to which it alters the intellectual structure of the state of the art. The structural variation approach focuses on the novel boundary-spanning connections introduced by a new article to the intellectual space. We validate the role of boundary-spanning in predicting future citations using three metrics of structural variation, namely, modularity change rate, cluster linkage, and centrality divergence, along with more commonly studied predictors of citations such as the number of co-authors, the number of cited references, and the number of pages. Main effects of these factors are estimated for five cases using zero-inflated negative binomial regression models of citation counts. Key findings indicate that 1) structural variations measured by cluster linkage are a better predictor of citation counts than the more commonly studied variables such as the number of references cited, 2) the number of coauthors and the number of references are both good predictors of global citation counts to a lesser extent, and 3) the centrality divergence metric is potentially valuable for detecting boundary-spanning activities at interdisciplinary levels. The structural variation approach offers a new way to monitor and discern the potential of newly published papers in context. The boundary-spanning mechanism offers a conceptually simplified and unifying explanation of the roles played by commonly studied extrinsic properties of a publication in the study of citation behavior.

Keywords: Boundary spanning, creativity, citation counts, structural variation, zero-inflated negative binomial regression

Introduction

A hallmark of scientific knowledge is its constant interplay with new ideas proposed by the scientific community. New ideas vary considerably in terms of how their values could be perceived. Some are warmly embraced upon their conception, whereas others may go through lengthy periods of uncertainty and controversy or even become totally ignored by the scientific community. A critical part of scientific inquiry is to discern where a new idea stands given what the scientific community knows as a whole. This is a cognitively demanding and conceptually challenging task. Not only do we need to have an up-to-date understanding of the intellectual structure of the relevant scientific fields, but also to identify exactly how a newly proposed idea is connected to the intellectual structure. New scientific publications arrive faster than what any individual can possibly read, analyze, and synthesize.

Detecting early signs of potentially valuable ideas has theoretical and practical implications. For instance, peer reviews of new manuscripts and new grant proposals are under a growing pressure of accountability for safeguarding the integrity of scientific knowledge and optimizing the allocation of limited resources (D. E. Chubin, 1994 ; Daryl E. Chubin & Hackett, 1990; Häyrynen, 2007; Hettich & Pazzani, 2006). Long-term strategic science and technology policies require visionary thinking and evidence-based foresights into the future (Cuhls, 2001; Martin, 2010; Miles, 2010). In foresight exercises on identifying future technology, experts' opinions were found to be overly optimistic on hindsight (Tichy, 2004). The increasing specialization in today's scientific community makes it unrealistic to expect an expert to have a comprehensive body of knowledge concerning multiple key aspects of a subject matter, especially in interdisciplinary research areas.

The value, or perceived value, of an idea can be quantified in many ways. For example, the value of a good idea can be measured by the number of people's life it has saved, the number of jobs it has created, or the amount of revenue it has generated. In the intellectual world, the value of a good idea can be measured by the number of other ideas it has inspired or the amount of attention it has drawn. In this article, we are concerned with identifying patterns and properties of information that can tell us something about the potential values of ideas expressed and embodied in scientific publications. A citation count of a scientific publication is the number of times other scientific publications have referenced to the publication. Using citations to guide the search for relevant scientific ideas by way of association, known as citation indexing, was pioneered by Eugene Garfield in the 1950s (Garfield, 1955). It is a general consensus that citation behavior can be motivated by both scientific and non-scientific reasons (Bornmann & Daniel, 2006). Citation counts have been used as an indicator of intellectual impact on subsequent research. There have been debates over the nature of citations and whether positive, negative, and self-citations should all be treated equally. Nevertheless, even a negative citation makes it clear that the referenced work cannot be simply ignored.

What do we know about factors that may influence citation counts in one way or another? One may address this question from a few different points of view, largely depending on where we draw our insights from, the past or the present. An article that has been highly cited so far is likely to remain highly cited according to the Matthew Effect (Merton, 1968). An article that has been frequently downloaded or viewed online is likely to become highly cited later on (Brody & Harnad, 2005; Kurtz et al., 2005). Relying on direct evidence such as visit counts, download counts, and citation counts that an article has already obtained has relatively lower risks than making assessments based on indirect evidence. The

downside of such approaches is that the analysis is not possible until a sufficient period of time elapses from the time of publication so that the article has a reasonable exposure to the scientific community. A much longer delay is expected to conduct a citation analysis because of the longer lifecycle of scholarly publication.

Researchers have searched for other clues that may inform us about the potential impact of a newly published scientific paper, especially clues that can be readily extracted from routinely available information at the time of publication instead of waiting for download and citation patterns to build up over time. Factors such as track record of authors, the prestige of authors' institutions, the prestige of the journal in which an article is published are among the most promising ones that can provide an assurance of the quality of the article to an extent (Boyack, Klavans, Ingwersen, & Larsen, 2005; Hirsch, 2007; Kostoff, 2007; van Dalen & Kenkens, 2005; Walters, 2006). The common assumption central to approaches in this category is that great researchers tend to continuously deliver great work and, along a similar vein, an article published in a high impact journal is also likely to be of high quality itself. On the one hand, these approaches avoid the reliance on data that may not be readily available upon the publication of an article and thus free analysts from constraints due to the lack of download and citation data. On the other hand, the sources of information used in these approaches are indirect to the new ideas reported in scientific publications. In an analogy, we give credits to an individual based on his/her credit history instead of assessing the risk of the current transaction directly. In such approaches, we will not be able to know where precisely the novelty of an idea is coming from. We will not be able to know whether similar ideas have been proposed in the past.

The approach that we will introduce in this article aims to provide specific trails of evidence to show why and how an idea is novel with reference to the current intellectual structure of a scientific domain. We conceptualize the development of scientific knowledge as a process of interplay between the intellectual structure and a stream of incoming new ideas conveyed in newly published scientific papers. Each new idea may alter the current intellectual structure or may leave the structure intact. The prediction of the potential value, or the impact, of an idea can be made computationally in terms of the degree of structural change introduced by the idea. In this context, we call this approach a structural variation model. For example, if a new idea connects previously disparate patches of knowledge, then its transformative potential is higher than the potential of ideas that are limited to well-trodden paths over the existing structure. The central idea of this approach is a boundary-spanning mechanism, which is conceptualized as a production rule that drives the citation process. The intellectual structure can be represented by networks of ideas. The conceptual change brought by newly published scientific articles to the intellectual structure can be quantified based on information that comes with the publication of such articles, notably the authorship and cited references. In order to validate that the structural variation model does capture insightful information about the potential value of an article, we investigate the extent to which structural variation measures of articles predict their subsequent citations along with other more commonly studied predictors such as the number of coauthors and the number of cited references. The structural variation model offers a conceptually simple and unifying explanation of several commonly identified citation predictors. For example, review and survey articles are often highly cited. An explanation in our model is that they tend to synthesize individual areas in a broader context than original research articles and they are more likely to include boundary spanning connections, which could be both intentionally and unintentionally when a large number of topics are reviewed. The number of coauthors has been

recognized as a potential factor to predict high citation counts. A possible explanation could be that multiple coauthors bring different areas of expertise and as a result, boundary spanning takes place as they collaborate.

In the following sections, we will describe the procedure of constructing baseline representations of the intellectual structure and define structural variation metrics. Then we will validate the role of structural variation mechanisms in predicting subsequent citations through a series of generalized linear models, which are particularly suitable for modeling count data with over dispersion and excessive zeros. These concepts will be discussed in more detail shortly. We will also demonstrate how the structural variation model provides a new way to interact and explore specific connections that make newly published scientific articles novel in the backdrop of the current intellectual structure.

The Structural Variation Model

There is a recurring theme from a diverse body of work on creativity. A major form of creative work is to bridge previously disjoint bodies of knowledge. Notable studies include the work of Ronald S. Burt in sociology (Burt, 2004), Donald Swanson in information science (Swanson, 1986a), and conceptual blending as a theoretical framework for exploring human information integration (Fauconnier & Turner, 1998). We have been developing an explanatory and computational theory of transformative discovery based on criteria derived from structural and temporal properties (Chen, 2011; Chen et al., 2009).

In the history of science, there are many examples of how new theories revolutionized the contemporary knowledge structure. For example, the 2005 Nobel Prize in medicine was awarded to the discovery of *Helicobacter pylori*, a bacterium which was not believed to be possible to find in human's gastric system (Chen, et al., 2009). In literature-based discovery, Swanson discovered previously unnoticed linkage between fish oil and Reynaud's syndrome (Swanson, 1986a). In terrorism research, before the September 11 terrorist attacks, it was widely believed that only those who directly witness a traumatic scene or directly experience a trauma could have the risk of post-traumatic stress disorder (PTSD); however, later research had shown that people may develop PTSD syndromes even by simply watching the coverage of a traumatic scene on TV (Chen, 2006). In drug discovery, one of the major challenges is to find new compound structures effectively in the vast chemical space that satisfy an array of constraints (Lipinski & Hopkins, 2004). In mapping scientific frontiers (Chen, 2003) and studies in science of science (Price, 1965), it would be particularly valuable if scientists, funding agencies, and policy makers can have tools that may assist them to assess the novelty of ideas in terms of their conceptual distance from the contemporary domain knowledge. In these and many more scenarios, a common challenge for coping with a constantly changing environment is to estimate the extent to which the structure of a network should be updated in respond to newly available information.

Many studies have addressed factors that could explain or even predict future citations of a scientific publication (Aksnes, 2003; Hirsch, 2007; Levitt & Thelwall, 2008; Persson, 2010). For example, is a paper's citation count last year a good predictor for new citations this year? Are the download times a good predictor of citations? Is it true that the more references a paper cites, the more citations it will receive later on? Similarly, the potential role of prestige, or the Matthew Effect coined by Robert Merton, has been commonly investigated, ranging from the prestige of authors to the prestige of journals in which articles are published (Dewett & Denisi, 2004). However, many of these factors are loosely and indirectly

coupled with the conceptual and semantic nature of the underlying subject matter of concern. We refer them as extrinsic factors in this article. In contrast, intrinsic factors have direct and profound connections with the intellectual content and structure. One example of intrinsic factor is concerned with the structural variation of a field of study. A notable example is the work by Swanson on linking previously disjoint bodies of knowledge, such as the connection between fish oil and Reynaud's syndrome (Swanson, 1986a).

Extrinsic Factors of Citations

Bornmann and Daniel (2006) reviewed 30 studies of citing behavior published between 1960s and mid-2005. The general tendency they found in the empirical studies is that citing behavior is motivated by many scientific and non-scientific factors. Their review summarizes the status of numerous detailed questions studied in the past. It has been noticed that some of the most highly cited articles are review articles or articles about methodology or tools (Tijssen, Visser, & van Leeuwen, 2002). Several studies have confirmed that the number of references cited by a paper appears to be a good predictor of its future citations among other factors such as the prestige of the paper's authors and its journal (Walters, 2006). For instance, review articles usually cite a large number of references and they tend to be highly cited (Aksnes, 2003). On the other hand, the number of references cited by an article alone is likely to be a poor proxy of the intellectual value of the article. It is reasonable to argue that which references an article cites matter more than how many references it cites. For example, if an article A cites considerably more references than another article B, then there are a few plausible explanations:

1. Article A addresses a topic much more extensively than article B does.
2. Article A, for example, a review paper, addresses more topics than article B does.
3. Article A, for example, a groundbreaking research paper, addresses a topic that builds on a synthesis of multiple topics, whereas article B synthesizes a less number of topics.

The list can go on further. The value of our approach is to take into account structural variations introduced by an article to provide additional insights into the relationship between an article and the state of the art.

Boyack et al. (2005) reported a method for predicting the importance of current papers based on journal importance, reference importance, and author reputation. Measures of importance in their study were based on citations to 780,049 papers published in 2002 with a window of citation from the beginning of 2002 till the end of 2003. 48% of papers were never cited in this window. The journal importance was calculated using the formula published by the Institute for Scientific Information (ISI). The reference importance was calculated as the number of times the reference in question was cited by papers published in 2002. The author reputation was calculated as the frequency of appearances of author-journal pairs in a four-year window before 2002. According to their report, a regression analysis of logarithmic transformed variables account for approximately 30% of variance. Among the three importance variables, the journal impact has the strongest correlation. They suggest properties such as these importance factors can be used to rank articles without the need to wait for citations to accumulate. While the ranking procedure they suggested is useful, their procedure implies an assumption that articles citing more popular references may get more citations. In this article, we take a different approach that focuses specifically on how the contemporary intellectual structure may change as a result of the way a new article links references. In other words, the argument that articles get more citations because they cite popular references only

reveals part of the story. Is it possible that frequent citations to an article are not because of the popularity of references it cites, but rather because of where these references are located in the intellectual structure of knowledge? Furthermore, is it possible to measure the value of a scientific contribution in terms of how many new ways of thinking it may introduce? This is indeed a major motivating question of our study.

Walters (2006) studied citations received by 428 articles published in 12 crime-psychology journals in a single year of 2003 and identified nine major predictors of citations, including author characteristics such as gender and citations to first authors' publications two years prior to the new publication, article characteristics such as co-authors, article length, and subject matter, and journal characteristics such as journal impact. Walters' study used negative binomial regression of the citations received by these 428 articles a year or two years after their publication. His study suggested that author impact may be a more powerful predictor of citations than the impact of a journal.

Kostoff (2007) compared highly and poorly cited research articles published in *The Lancet* and found that the most cited articles tend to have more coauthors, cite more references, have a longer abstract and more pages. Highly cited papers tend to report clinical trials of much larger sample sizes than poorly cited papers. In a specific context, the h-index was found to have a strong correlation with the number of citations – the h-index calculated based on the first 12 years in a sample data has a correlation coefficient of 0.60 with citations in the second 12 years (Hirsch, 2007).

Levitt and Thelwall (2008) studied highly cited articles in six subject areas and found that predicting citation ranking of highly cited articles using the subtotals of citations in the 5^{th} and 6^{th} year turned out to be more accurate than using the total of citations in the first six years. Skilton (2009) found frequently cited coauthors and authors with diverse disciplinary backgrounds tend to be highly cited in natural sciences. In contrast, the variety of disciplines represented collectively by the authors of an article does not influence citations. Persson (2010) addressed the question whether highly cited papers are more international, i.e. written by coauthors from different geographic locations, than less highly cited ones. Based on data from four research areas, from three universities, four cities and two countries, he concluded that international papers dominate highly cited papers from small countries, but not well represented in high impact papers overall.

In addition to characteristics of authors and their institutions, features of articles such as the length of title, number of figures and tables, and number and recency of references and research methodology have been studied in the literature (Haslam et al., 2008).

Intrinsic Factors of Citations

Researchers have made various attempts to characterize future citations and identify emerging core articles (Shibata, Kajikawa, & Matsushima, 2007; Walters, 2006). Shibata et al., for example, studied citation networks in two subject areas, Gallium Nitride and Complex Networks, and found that while past citations are a good predictor of near-future citations, the betweenness centrality is correlated with citations in a longer term.

Upham, Rosenkopf, and Ungar (2010) studied the role of cohesive intellectual communities – schools of thoughts – in promoting and constraining knowledge creation. They analyzed publications on management and concluded that it is significantly beneficial for new knowledge to be a part of a school of thought and the most influential position within a school of thought is in the semi-periphery of the school.

In particular, boundary-spanning research positioned at the semi-periphery of a school would attract attention from other schools of thought and receive the most citations overall. Their study used a zero-inflated negative binomial regression (ZINB). Negative binomial regression models have been used to predict the expected mean patent citations (Fleming & Bromiley, 2000). Hsieh (2011) studied inventions as a combination of technological features. In particular, the closeness of features plays an interesting role. Neither overly related nor loosely related features are good candidates for new inventions. Useful inventions arise with rightly positioned features where the cost of synthesis is minimized.

Takeda and Kajikawa (2010) reported three stages of clustering in citation networks. In the first stage, core clusters are formed, followed by the formation of peripheral clusters and the continuous growth of the core clusters. Finally, the core clusters' growth becomes predominant again. Buter, Noyons, and van Raan (2011) studied the emergence of an interdisciplinary research area from fields that did not show interdisciplinary connections before. They used journal subject categories as a proxy for fields and citations as a measure of interdisciplinary connection.

Lahiri et al. addressed how structural changes of a network may influence the spread of information over the network (Lahiri, Maiya, Sulo, Habiba, & Wolf, 2008). Although they did not study bibliographic networks per se, their study indicates predictions made about how information spreads over a network are sensitive to structural changes of the network. This observation underlines the importance of taking structural change into account in the development of metrics based on topological properties of networks.

Leydesdorff (2001) raised questions (p. 146) that are closely related to what we are addressing: "How does the new text link up to the literature, and what is its impact on the network of previously existing relations?" He took a quite different approach and analyzed word occurrences in scientific papers from an information-theoretic perspective. In his approach, the publication of a paper is perceived as an event that may lead to the reduction of uncertainty involved in the current state of knowledge. He devised diagrams that depict pathways of how a particular paper improves the efficiency of communication. Although the information-theoretic approach and our structural variation approach currently operate on different units of analysis with distinct theoretical underpinnings, both share the fundamental concern of changes introduced by newly published scientific papers on the existing body of knowledge.

As shown above, many studies in the literature have addressed factors that may influence citations. The value of our work is the introduction of the structural variation paradigm along with computational metrics that can be integrated into interactive exploration systems to better understand precisely the impact of individual links made by a new article.

In this article, we conceptualize that structural variation is an essential process that advances scientific knowledge. The intellectual structure at a given point of time is subject to structural changes introduced by newly published scientific articles. The intellectual structure may be represented in many different types of networks, including networks of co-cited references, networks of co-cited authors, or networks of co-occurring keywords. Given a scientific article, publications prior to the publication of the article form a baseline representation of the intellectual structure at the specific time of publication. Structural variation metrics measure the degree of structural change introduced by information derived from the article. An overview of the procedure is depicted in Figure 1. Individual steps are explained in corresponding sections below.

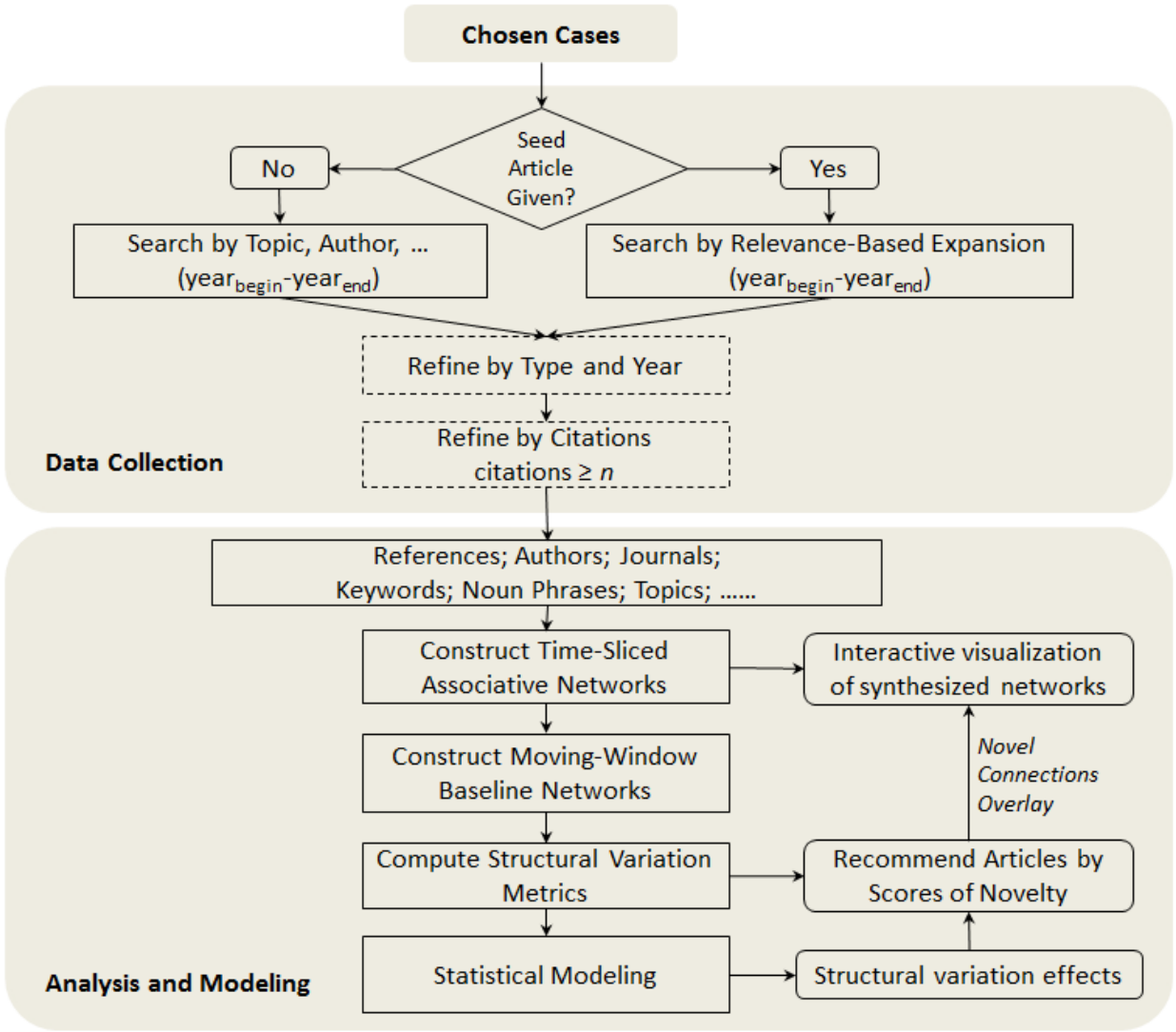

Figure 1. An overview of the structural variation model.

Data Collection

First, we choose a topic or a target field of study. In this article, we include five cases drawn from four research areas such as terrorism, mass extinction, complex network analysis, and knowledge domain visualization. We have studied some of these research areas in our previous studies. Choosing a familiar domain can help us to concentrate on new insights to be revealed by our new approach. We also include a dataset of articles defined by the journals in which they are published. We use the Web of Science as the source of our data, although the method is applicable to other sources of citation data such as Scopus and Google Scholar. In the rest of this article, bibliographic records are retrieved from the Web of Science unless stated otherwise.

Given a chosen topic, a dataset can be constructed in two ways. One is to search for records of articles by relevance determined by matching terms and other attributes. For instance, a dataset of bibliographic records on terrorism research can be constructed by using a topic search for the occurrence of the term

terrorism in the title, the abstract, or the keyword field of a record. It is possible to define a dataset based on specific source journals, for example, all articles published in *Nature* and *Science*. The other is to locate articles by association formed by cited references. The second method starts with a seed article. All articles that have at least one overlapping reference with the seed article are identified as related records. The global citation count of an article is the total number of times the article is referenced by other articles in the Web of Science at the time of retrieval. The local citation count is the total number of times the article is referenced by other articles within the retrieved dataset, which is a subset of the entire collection of the Web of Science. The global citation count is used in the analysis reported in this article because we assume that it is a less biased proxy of impact than the local citation count.

A seed article is used as a starting point to reconstruct the structural representations prior to and after the publication of the seed article. We use a two-stage expansion process. First, all the references cited by the seed article *s* are identified in the Web of Science as a set *CitedBy*(*s*). Second, articles in the Web of Science that cite at least one reference in *CitedBy* (*s*) form a set $S_{Expanded}(s)$.

$$S_{Expanded}(s \in Articles) = \{a \in Articles | CitedBy(a) \cap CitedBy(s) \neq \emptyset\}$$

The final set of articles serves as the representation of the population of the original seed article. This process of citation expansion can be repeated as many times as meaningful to the purpose of the intended study, for example, $\bigcup_{a \in S_{Expanded}(s)} S_{Expanded}(a)$. A description of the process is detailed in our earlier publication on forward expansion (Chen, Lin, & Zhu, 2006). Although we do not rule out the possibility that a relevant article does not have any overlap with an expanded set of articles generated in this procedure, for example, the relevance is purely established at a semantic level, in practice it would be rare to come across an article that is relevant but it cites a totally different body of literature, especially when the body of literature is identified by the two-step citation expansion. Such articles would either ignore the relevant literature or completely innovative. Two cases are based on seeded datasets, namely, Complex Network Analysis and CiteSpace Expanded, whereas the other three cases are based on the relevance of topic search.

In order to explore various configurations of the structural variation model, we apply the procedure to a number of datasets representing a diverse range of topics and subject domains. The background of each case is outlined as follows. Additional references are provided for further reading.

Complex Network Analysis

This dataset overlaps with the Small-World Networks dataset, but the two datasets were constructed differently. A topic search for complex network analysis revealed two most frequently cited articles (Barabasi & Albert, 1999) and (Watts & Strogatz, 1998). We constructed an updated dataset with these two articles as seed articles. First, one seed article was used to form a sub-dataset. Then, the two datasets were merged to form a combined dataset on complex network analysis.

The seed article (Barabasi & Albert, 1999) has the highest citation count. As of August 5, 2011, it has been cited 5,792 times in the Web of Science. The article (Watts & Strogatz, 1998) has the second highest citation count of 5,291. For each of the seed articles, we retrieved all the records that share at least one cited reference with the seed article. Then we merged all the records and formed a combined dataset. The seed Barabasi_1999 has 8,919 related articles published between 1980 and 2011. If we limited to the

period of 1996-2004 and to original research articles, reviews, and proceedings, 2,326 records were retained in the subset. The seed article Watts_1998 has 14,393 related records. Interestingly, this is 1.6 times more than what Barabasi_1999 has, although Barabasi_1999 has 500 more citations. It appears that Watts_1998 involves a wider range of topics than the Barabasi_1999 article. The merged dataset from the two subsets contains 6,764 records. We first analyzed the dataset seeded by Barabasi1999 and then the dataset seeded by Watts1998, and finally, the merged dataset.

Mass Extinctions

The Mass Extinctions research is not new. Some of its major groundbreaking works appeared in early 1980s. On the other hand, the field is as active as ever because many questions remain concerning mass extinctions. Our previous study of the evolution of mass extinctions research between 1981 and 2004 revealed a shift of research focus in early 1990s from the K-T boundary mass extinction to the Permian extinction (Chen, 2006; Chen, Cribbin, Macredie, & Morar, 2002). We expect to detect considerable structural variations within the range of 1991-2010.

The Mass Extinctions dataset contains 1,745 records of journal articles, conference proceedings, and review articles published between 1991 and 2010 on mass extinctions. Records were retrieved using a topic search query "mass extinction" on April 2010. The dataset does not cover articles published later in the year 2010, but this does not affect our analysis because we take the exposure time into account in generalized linear models.

Terrorism

This dataset consists of bibliographic records resulted from a topic search in the Web of Science on terrorism between 1996 and 2005. A total of 1,303 articles were retrieved on July 11, 2011. The citation counts of articles in the dataset indicate the number of citations an article has received since its publication up to July 2011. For example, an article published in January 1996 would have the longest exposure time of over 15 years, whereas an article published in December 2005 would have the shortest exposure time of more than 5 years. The exposure time is taken into account in the construction of generalized linear models. We will provide more details on this shortly.

In a previous study, we analyzed an earlier period of the same subject (1990-2003) (Chen, 2006). Several conceptual transformations were revealed by our previous study. In particular, we were able to identify a conceptual change from the topological properties of a synthesized network of co-cited references. Prior to the terrorist attacks on September 11, 2001, it was generally believed that post-traumatic stress disorder (PTSD) would only be found in people who had a direct experience of trauma. Researchers then discovered after the extensive coverage of the September 11 terrorist attacks on mass media people who were far away from New York could also develop PTSD symptoms (Galea et al., 2002). This is one example of boundary spanning in that a new idea redefines the boundary of a topic area and introduces fundamental changes to the original intellectual structure. The terrorist attacks on September 11, 2001 occurred within the 1996-2005 range. Therefore we expect that the updated new dataset would provide relevant information for us to study the dynamics of the structural variation.

Two cases are derived from this dataset with different configurations of the length of time slices. In one case, we use three-year time slices and in another case, we use two-year time slices in order to identify the procedural implications of using time slices of different duration.

CiteSpace Expanded

The case is seeded by our 2006 article on CiteSpace (Chen, 2006). This dataset consists of papers that have at least one cited reference in common with our seed article. Based on the content of the seed article, we anticipate that there will be three sub-areas involved in the dataset, namely scientometrics, mass extinctions, and terrorism. CiteSpace is a scientometric tool that has been applied to the analysis of mass extinctions and terrorism.

Baseline Networks

The basic assumption in the structural variation approach is that the extent of a departure from the current intellectual structure is a necessary condition for a potentially transformative idea in science. In other words, a potentially transformative idea needs to bring changes to the existing structure of knowledge in the first place. In order to measure the degree of structural variation introduced by a scientific article, the intellectual structure at a particular moment of time needs to be represented in such a way that structural changes can be computationally detected and manually verifiable. Bibliographic networks can be computationally derived from scientific publications. Research in scientometrics and citation analysis routinely uses citation and co-citation networks as a proxy of the underlying intellectual structure. In this article, we focus on using several types of co-citation and co-occurrence networks as the representation of a baseline network.

A network represents how a set of entities are connected. Entities are represented as nodes, or vertices, in the network. Their connections are represented as links, or edges. Relevant entities in our context include several types of information that can be computationally extracted from a scientific article, such as references cited by the article, authors and their affiliations, the journal in which the article is published, and keywords in the article. In this article, we limit our discussions to networks that are formed with a single type of entities, although networks of multiple types of entities are worth considering once we establish a basic understanding of structural variations in networks of a single type of entities.

Once the type of entities is chosen, the nature of the interconnectivity between entities is to be specified to form a network. Networks of co-occurring entities represent a wide variety of types of connectivity. A network of co-occurring words represents how words are related in terms of whether and how often they appear in the vicinity of each other. Co-citation networks of entities such as references, authors, and journals can be seen as a special case of co-occurring networks. For example, co-citation networks of references are networks of references that appear together in the bodies of scientific papers – these references are co-cited.

Networks of co-cited references represent more specific information than networks of co-cited authors because references of different articles by the same author would be lumped together in a network of co-cited authors. Similarly, networks of co-cited references are more specific than networks of co-cited journals. We refer such differences in specificity as the granularity of networks. Measurements of structural variation need to take the granularity factor into account because it is reasonable to expect that networks at different levels of granularity would lead to different measures of structural variations.

Another decision to be made about a baseline network is a sampling issue. Taking a particular year as a standing point to look at in the past, how far back should we consider in the construction of a baseline network that would adequately represent the underlying intellectual structure? Does the network become

more accurate if we go back more into the past? Will it be more efficient if we limit it to the most recent years that really matter the most? In this article, given articles published in a particular year *Y*, the baseline network represents the intellectual structure using information from articles published up to year *Y-1*. Two types of baseline networks are investigated in this article: ones using a moving window of a fixed size [*Y-k, Y-1*] and ones using the entire history (Y_0, *Y-1*], where Y_0 is the earliest year of publication for records in the given dataset.

Structural Variation Metrics

We expect that the degree of structural variation introduced by a new article can offer prospective information because of the boundary spanning mechanism. If an article introduces novel links that span the boundaries of different topics, then we expect this signifies its potential in taking the intellectual structure for a new turn.

Given a baseline network, structural variations can be measured based on information provided by a particular article. In this article, we introduce three metrics of structural variation. Each metric quantifies the degree of change in the baseline network introduced by information provided by an article. No usage data is involved in the measurement. The three metrics are modularity change rate, inter-cluster linkage, and centrality divergence. The definitions of the first two metrics depend on a partition of the baseline network, but the third one does not. A partition of a network decomposes the network into non-overlapping groups of nodes. For example, clustering algorithms such as spectral clustering can be used to partition a network.

Principles

The theoretical underpinning of the structural variation is that scientific discoveries, at least a subset of them, can be explained in terms of boundary spanning, brokerage, and synthesis mechanisms in an intellectual space (Chen, et al., 2009). This conceptualization generalizes the principle of literature-based discovery pioneered by Swanson (Swanson, 1986a, 1986b), which assumes that connections between previously disparate bodies of knowledge are potentially valuable. In Swanson's famous ABC model, the relationships AB and BC are known in the literature. The potential relationship AC becomes a candidate that is subject to further scientific investigation (Weeber, 2003). Our conceptualization is more generic in several ways. First, in the ABC model, the AC relation changes an indirect connection to a direct connection, whereas our structural variation model makes no assumption about any prior relations at all. Second, in the ABC model, the scope of consideration is limited to relationships involving three entities. In contrast, our structural variation model takes a wider context into consideration and addresses the novelty of a connection that links groups of entities as well as connections linking individual entities. Because of the broadened scope of consideration, it becomes possible to search for candidate connections more effectively. In other words, given a set of entities, the size of the search space of potential connections can be substantially reduced if additional constraints are applicable for the selection of candidate connections. For example, the structural hole theory developed in social network analysis emphasizes the special potential of nodes that are strategically positioned to form brokerage, or boundary spanning, links and create good ideas (Burt, 2004; Chen, et al., 2009).

Modularity Change Rate (MCR)

Given a partition of a network, i.e. a configuration of clusters, the modularity of the network measures the degree of interconnectivity among the groups of nodes identified by the partition. If different clusters are loosely connected, then the overall modularity would be high. In contrast, if clusters are interwoven, then the modularity would be low. We follow Newman's algorithm (Newman, 2006) to calculate the modularity with reference to a cluster configuration generated by spectral clustering (Chen, Ibekwe-SanJuan, & Hou, 2010; Luxburg, 2006). Suppose the network G is partitioned by a partition C into k clusters such that $G = c_1 + c_2 + \ldots + c_k$, $Q(G)$ is defined as follows, where m is the total number of edges in the network G. n is the number of nodes in G. $\delta(c_i, c_j)$ is known as the Kronecker's delta. It is 1 if nodes n_i and n_j belong to the same cluster and 0 otherwise. $deg(n_i)$ is the degree of node n_i. The range of $Q(G)$ is between -1 and 1.

$$Q(G, C) = \frac{1}{2m} \sum_{i,j=0}^{n} \delta(c_i, c_j) \cdot (A_{ij} - \frac{\deg(n_i) \cdot \deg(n_j)}{2m})$$

The modularity of a network is a measure of the overall structure of the network. Its range is between -1 and 1. The Modularity Change Rate of a scientific paper measures the relative structural change due to the information from the published paper with reference to a baseline network. For each article a, and a baseline network $G_{baseline}$, we define the Modularity Change Rate (MCR) as follows:

$$MCR(a) = \frac{Q(G_{baseline}, C) - Q(G_{baseline} \oplus G_a, C)}{Q(G_{baseline}, C)} \cdot 100$$

where $G_{baseline} \oplus G_a$ is the updated baseline network by information from the article a. For example, suppose reference nodes n_i and n_j are not connected in a baseline network of co-cited references but they are co-cited by article a, a new link between n_i and n_j will be added to the baseline network. In this way, the article changes the structure of the baseline network.

Intuitively, adding a new link anywhere in a network should not increase the modularity of the network. It should either reduce it or leave it intact. However, the change of modularity is not a monotonic function as we initially expect. In fact, it depends on where the new link is added and how the network is structured. Adding a link may reduce the proportion of the modularity in some clusters, but it may increase the modularity in other clusters in the network. Thus, the overall modularity change is not monotonic. Diagrams in the following figure illustrate these scenarios.

Without losing any generality, assume that an article adds one link at a time to a given baseline network. If the new link connects two distinct clusters, then it has no effect on the corresponding term in the updated modularity because by definition $\delta_{ij}=0$ and the corresponding term becomes 0. Such a ink is illustrated by the dashed link $e_{5,10}$ in the top diagram in Figure 2. The new link e_{ij} will increase the degree of nodes i and j by one, i.e. $deg(i)$ will become $deg(i)+1$. The total number of edges m will increase to $m + 1$. A simple calculation at the bottom of Figure 2 shows that terms in the modularity formula involving blue links will decrease from their previous values. However, if the network has clusters such as C_A with no changes in node degrees, then the corresponding values of terms of lines in red will increase from their previous values as the denominator increases from $2m$ to $2(m + 1)$. In summary, the updated modularity may increase as well as decrease, depending on the structure of the network and where the new link is

added. With this particular definition of modularity, between-cluster links are always associated with a zero valued term in the overall modularity formula due to the Kronecker's delta. What we see in the change of modularity is a combination of results from several scenarios that are indirectly affected by the newly added link. We will introduce our next metric to reflect the changes in terms of between-cluster links directly.

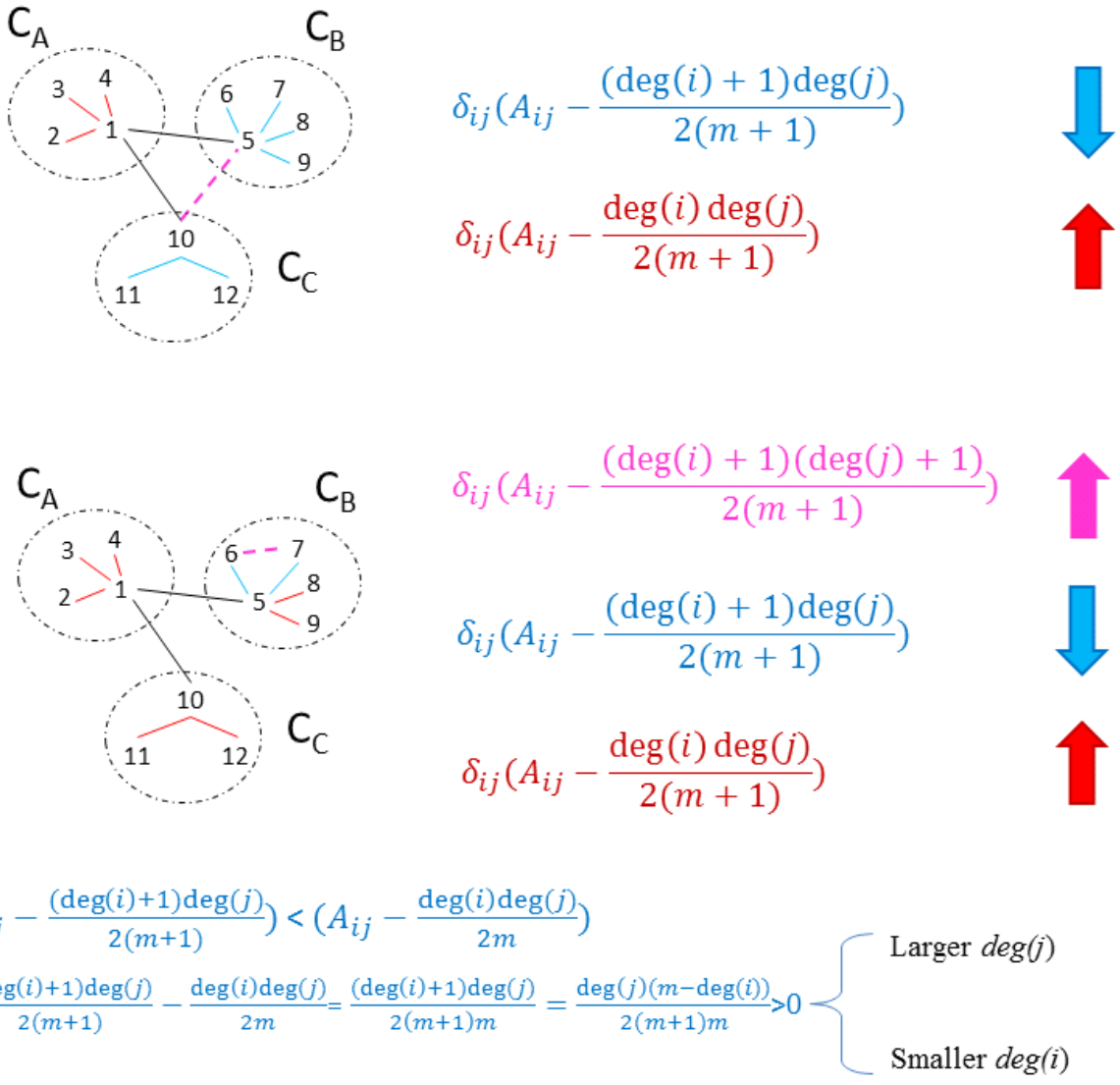

Figure 2. Scenarios that may increase or decrease individual terms in the modularity metric.

As shown in Figure 3, two articles with the highest modularity change rates in a small dataset of papers that cited (Chen, 2006) were review articles. Review articles tend to cover many topics, as one would expect, which would mean adding many boundary spanning co-citation connections to baseline networks and thus lead to a substantial structural variation in terms of modularity. The two review papers have different distributions of boundary spanning links: one with references concentrated on a small number of

clusters in recent years and the other with references traced back several decades ago in distinct clusters. Both papers cited the informetrics cluster (#11).

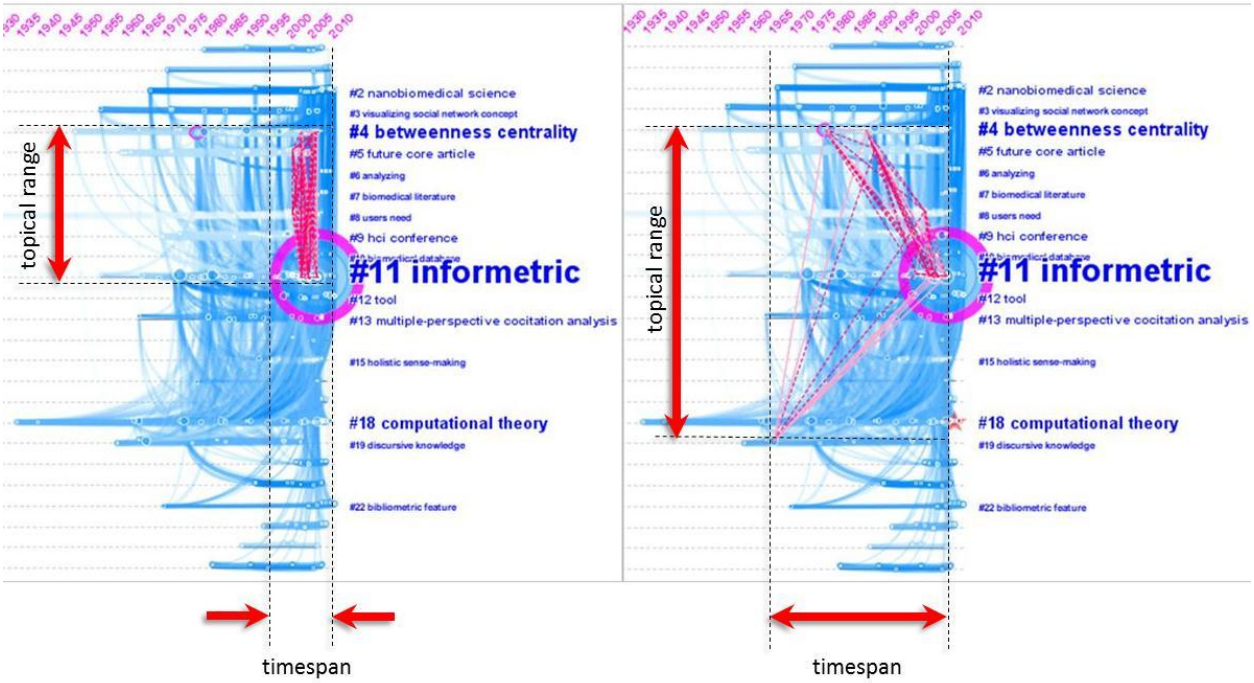

Figure 3. The depth and breadth of novel connections made by two articles with the highest modularity change rate. Both are review articles. Barilan 2008 (Left) cited references in recent years within a relatively small number of clusters, whereas Morris 2008 (Right) cited many earlier publications across a wider range of clusters.

Cluster Linkage (CL)

The Cluster Linkage (CL) measures the overall structural change introduced by an article *a* in terms of new connections added between clusters. Its definition assumes a partition of the network. We introduce a function of edges $\lambda(c_i,c_j)$ which is the opposite of δ_{ij} used in the modularity definition. The value of λ_{ij} is 1 for an edge across distinct clusters c_i and c_j. It will be 0 for edges within a cluster. λ_{ij} will allow us to concentrate on between-cluster links and ignore within-cluster links, which is the opposite of how the modularity metric is defined. The new metric *Linkage* is the sum of all the weights of between-cluster links e_{ij} divided by K - the total number of clusters in the network. Linking to itself is not allowed, i.e. we assume e_{ii} =0 for all nodes. Using link weights makes the metric sensitive to links that strengthen connections between clusters in addition to novel links that make unprecedented connections between clusters.

It is possible to take into account the size of clusters that a link is connecting so that connections between larger-sized clusters become more prominent in the measurement. For example, one option is to multiple each e_{ij} by $\sqrt{size(c_i) \cdot size(c_j)}/\max(size(c_k))$. In this article, the metric is defined without such modifications for simplicity. Suppose C is a partition of G, the *Linkage* metric is defined as follows:

$$Linkage(G,C) = \frac{\sum_{i \neq j}^{n} \lambda_{ij} \boldsymbol{e_{ij}}}{K}$$

$$\lambda_{ij} = \begin{cases} 0, n_i \in c_j \\ 1, n_i \notin c_j \end{cases}$$

The *Cluster Linkage* is defined as the difference of *Linkage* before and after new between-clusters links added by an article *a*.

$$CL(a) = \Delta Linkage(a) = Linkage(G_{baseline} \oplus G_a, C) - Linkage(G_{baseline}, C)$$

Linkage(G + ΔG) is always greater than or equal to *Linkage*(G). Thus, CL is non-negative.

Centrality Divergence (C_{KL})

The Centrality Divergence metric measures the structural variation caused by an article *a* in terms of the divergence of the distribution of betweenness centrality $C_B(v_i)$ of nodes v_i in the baseline network. This definition does not involve any partitions of the network. If *n* is the total number of nodes. The degree of structural change $C_{KL}(G, a)$ can be defined in terms of the K-L divergence.

$$C_{KL}(G_{baseline}, a) = \sum_{i=0}^{n} p_i \cdot \log\left(\frac{p_i}{q_i}\right)$$

$$p_i = C_B(v_i, G_{baseline})$$

$$q_i = C_B(v_i, G_{updated})$$

For nodes where p_i=0 or q_i=0, we reset them as a small number 10^{-6} to avoid *log*(0).

Statistical Models

We constructed negative binomial (NB) and zero-inflated negative binomial (ZINB) models to validate the role of structural variation in predicting future citation counts of scientific publications. The negative binomial distribution is generated by a sequence of independent Bernoulli trials. Each trial is either a 'success' with a probability of *p* or a 'failure' with a probability of (1-*p*). Here the terminology of success and failure in this context does not necessarily represent any practical preferences. The random number of successes ***X*** before encountering a predefined number of failures ***r*** has a negative binomial distribution:

$$\boldsymbol{X} \sim \mathrm{NB}(r, p)$$

One can adapt this definition to describe a wide variety of count events. Citation counts belong to a type of count events with an over-dispersion, i.e. the variance is greater than the mean. NB models are commonly used in the literature to study this type of count events. Two types of dispersion parameters are used in the literature, θ and α, where θ•α=1.

Zero-inflated count models are commonly used to account for excessive zero counts (Hilbe, 2011; Lambert, 1992). Zero-inflated models include two sources of zero citations: the point mass at zero $I_{\{0\}}(y)$ and the count component with a count distribution f_{count}(counts) such as negative binomial or Poisson (Zeileis, Kleiber, & Jackman, 2011). The probability of observing a zero count is inflated with probability $\pi = f_{zero}$(zero citations).

$$f_{zero\text{-}inflated}(\text{citations}) = \pi \times I_{\{0\}}(\text{citations}) + (1- \pi) \times f_{count}(\text{citations})$$

ZINB models are increasingly used in the literature to model excessive occurrences of zero citations (Fleming & Bromiley, 2000; Upham, et al., 2010). The report of a ZINB model consists of two parts: the count model and the zero-inflated model. One way to test whether a ZINB model is superior to a corresponding NB model is known as the Vuong test. The Vuong test is designed to test the null hypothesis that the two models are indistinguishable. Akaike's Information Criterion (AIC) is also commonly used to evaluate the goodness of a model. Models with lower AIC scores are regarded as better models in terms of the relative goodness of fit.

In this article, we focus on global citation counts of scientific publications recorded in the Web of Science. NB models are defined as follows using log as the link function.

Global citations ~ Coauthors + Modularity Change Rate + Cluster Linkage + Centrality Divergence + References + Pages

Global citations is the dependent variable. *Coauthors* is a factor of three levels of 1, 2, and 3. Level 3 is assigned to articles with 3 or more coauthors. *Coauthors* is an indirect indicator of the extent to which an article synthesizes ideas from different areas of expertise represented by each coauthor.

Three structural variation metrics are included as co-variants in generalized linear models, namely *Modularity Change Rate* (MCR), *Cluster Linkage* (CL), and *Centrality Divergence* (C_{KL}). According to our theory of creativity, groundbreaking ideas are expected to cause strong structural variations. If global citation counts provide a reasonable proxy of recognitions of intellectual contributions in a scientific community, we would expect that at least some of the structural variation metrics will have statistically significant main effects on global citations.

The number of cited references and the number of pages are commonly reported in the literature as good predictors of citations. In order to compare the effects of structural variation with these commonly reported extrinsic properties of scientific publications, *References* and *Pages* are included in the models. Our theory offers a simpler explanation why the more references a paper cites, the more citations it appears to get. Due to the boundary spanning synthetic mechanism, an article needs to explain multiple parts and how they can be innovatively connected. This process will result in citing more references than an article that covers a narrower range of topics. Review papers by their nature belong to this category.

It is known that articles published earlier tend to have more citations than articles published later. The exposure time of an article is included in the NB models in terms of a logarithmically transformed year of publication of an article.

An intuitive way to interpret coefficients in NB models is to use incidence rate ratios (IRRs) estimated by the models. For example, if *Coauthors* has an IRR of 1.5, it means that as the number of coauthors increases by one the global citation counts would be expected to increase a factor of 1.5, i.e. increasing 1.5 times, while holding other variables in the model constant. In our models, we will particularly examine statistically significant IRRs of structural variation models.

Zero-inflated negative binomial models (ZINB) use the same set of variables. The count model of ZINB is identical to the NB model described above. The zero-inflated model of ZINB uses the same set of variables to predict the excessive zeros. We found little in the literature about good predictors of zeros in a comparable context. We choose to include all the six variables in the zero-inflated model to provide a broader view of the zero-generating process. ZINBs are defined as follows:

Global citations ~ Coauthors + Modularity Change Rate + Cluster Linkage + Centrality Divergence + References + Pages

Zero citations ~ Coauthors + Modularity Change Rate + Cluster Linkage + Centrality Divergence + References + Pages

Results

Unless stated otherwise, baseline networks are formed with a 2-year sliding window [Y-2, Y-1] for papers published in year Y. Co-citation links made by top 100 most cited articles in each year of the sliding window are used to construct a baseline network.

Complex Network Analysis (1996-2004)

Figure 4 shows an overview of a network of co-cited references in this field. Only major clusters' labels are shown in the figure. The largest cluster (Cluster #8) is labeled as complex network. The groups of clusters in the lower left of the visualization include social capital (Cluster #12), and old helping partner (Cluster #11). The groups of clusters in the lower right of the visualization include striate cortical activity (Cluster #2), stochastic resonance (Cluster #1), and phase synchronization (Cluster #5). Dashed lines in red are novel connections made by (Watts & Strogatz, 1998) at the time of its publication. The article has the highest scores in Cluster Linkage and C_{KL} scores, 5.43 and 1.14, respectively. The figure offers a visual confirmation that the article was indeed making boundary-spanning connections. Recall that the data set was constructed by expanding the seed article based on forward citation links. These boundary-spanning links provide empirical evidence that the groundbreaking paper was connecting two groups of clusters. The emergence of Cluster #8 complex network was the consequence of the impact.

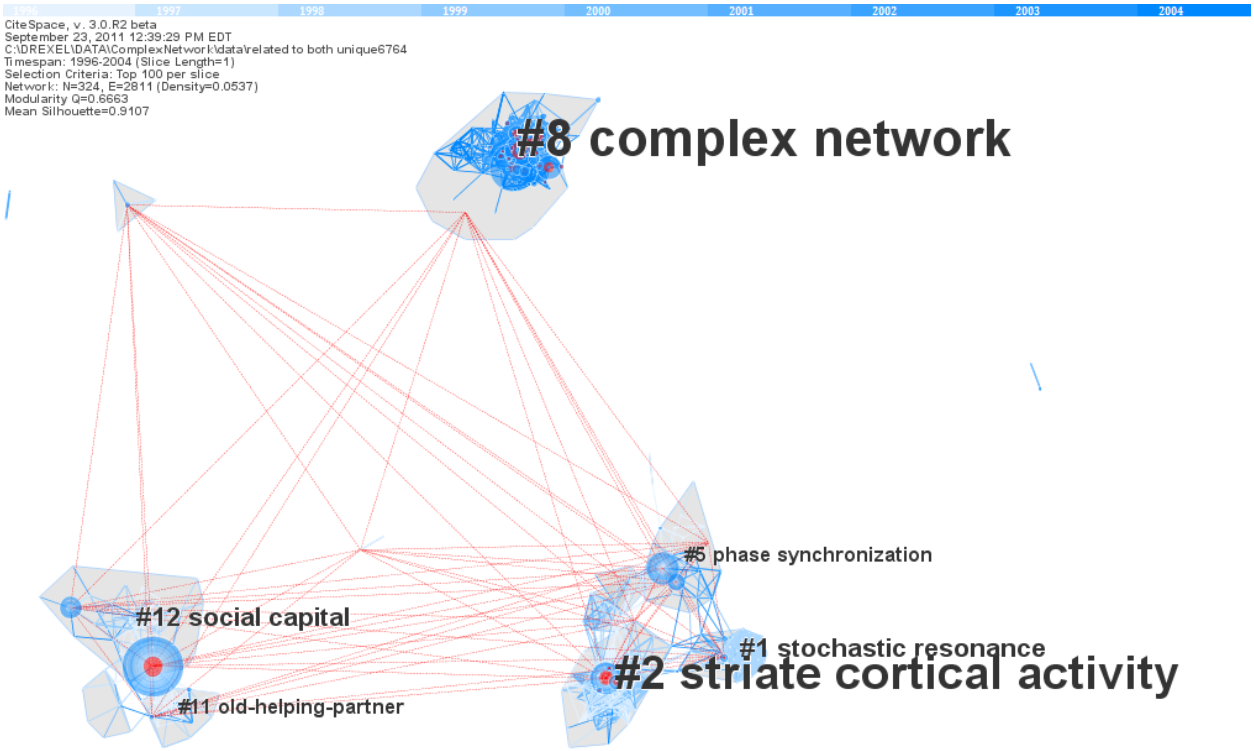

Figure 4. A network of co-cited references derived from the Complex Network Analysis (1996-2004). Dashed lines in red are novel connections made by the groundbreaking article (Watts & Strogatz, 1998), which has the highest scores in both Cluster Linkage (5.43) and Centrality Divergence (1.14).

Figure 5 is a close-up view of the lower right region shown in Figure 4. Dashed lines in red depict the novel links made by Watts and Strogatz in 1998. The references connected by the new links are labeled.

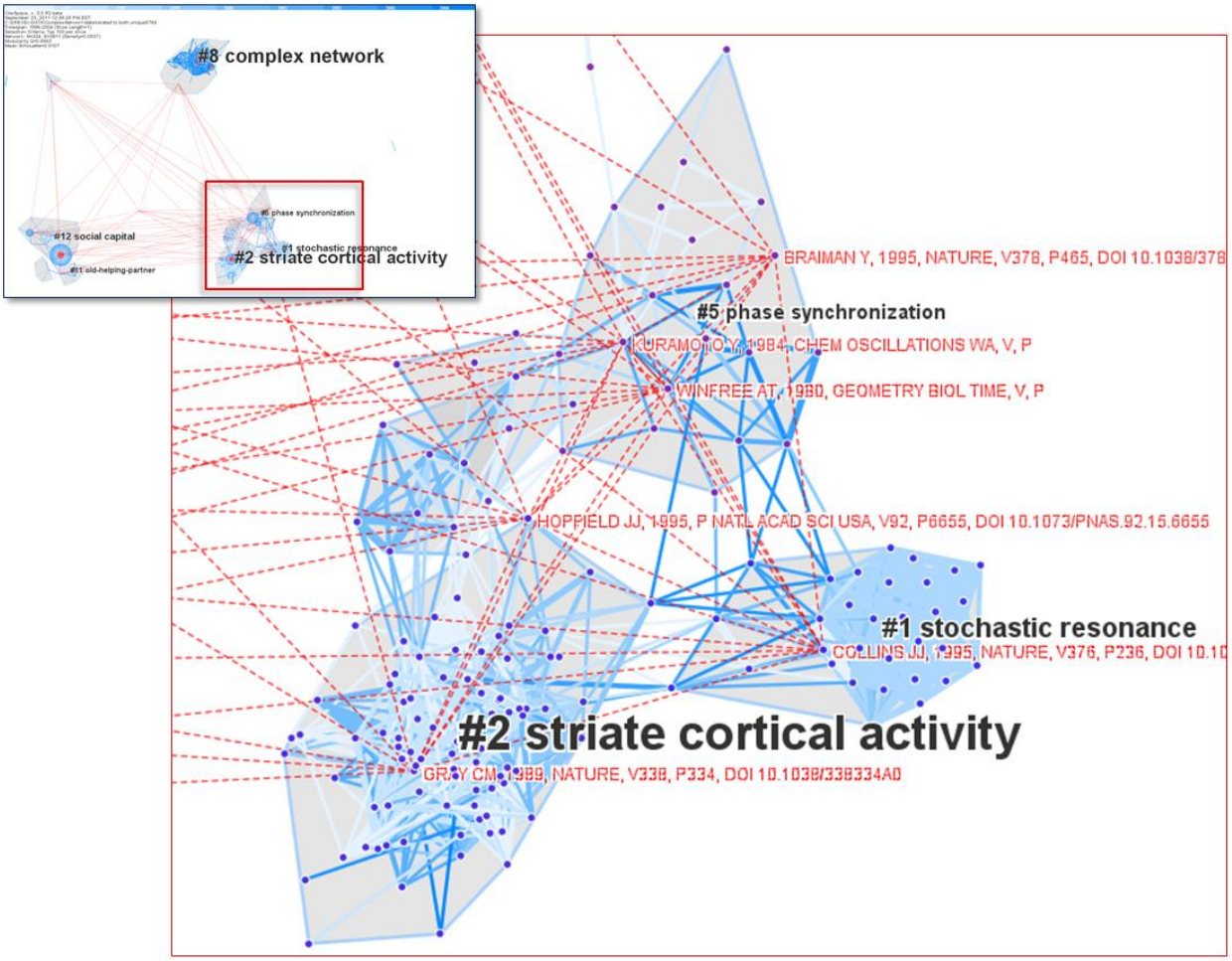

Figure 5. A close-up view of the lower right region shown in Figure 4, showing novel links between distinct clusters of co-cited references.

Table 1 summarizes the results of five NB regression models with different types of networks. They have an average dispersion parameter θ of 0.5270, which is equivalent to an alpha of 1.8975. *Coauthors* has an average IRR of 1.3278. *References* has an average IRR of 1.0126. *Pages* has an average IRR of 0.9714. The effects of the three variables are consistent and stable across the five types of networks. In contrast, the effects of structural variations are less stable. On the other hand, structural variations appear to have a stronger impact on global citations than other more commonly studied measures such as *Coauthors* and *References*. For example, CL has an IRR of 3.160 in networks of co-cited references and an IRR of 1.33×10^{8} in networks of noun phrases. IRRs that are greater than 1.0 predict an increase of global citations.

Table 1. Negative binomial regression models (NBs) of Complex Network Analysis (1996-2004) at five different levels of granularity of units of analysis. References involves the least amount of ambiguity with the finest granularity, whereas the other four types of units introduce ambiguity at various levels. Models constructed with units of higher ambiguity are slightly improved in terms of Akaike's Information Criterion (AIC).

Data Source: Complex Network Analysis (1996-2004), top100 records per time slice, 2-year sliding window										
Unit of Analysis	Reference		Keyword		Noun Phrase		Author		Journal	
Relation	Co-citation		Co-occurrence		Co-occurrence		Co-citation		Co-citation	
Offset (Exposure)	$\log_2$(Year)		$\log_2$(Year)		$\log_2$(Year)		$\log_2$(Year)		$\log_2$(Year)	
Number of Citing Articles	3,515		3,072		3,254		3,271		3,271	
Global Citations	**Incidence Rate Ratios (IRRs) in NB Models**									
Coauthors	**1.306**	**.000**	**1.298**	**.000**	**1.326**	**.000**	**1.359**	**.000**	**1.350**	**.000**
Modularity Change Rate	**1.083**	**.025**	1.038	.086	1.047	.305	1.055	.276	1.060	.180
Weighted Cluster Linkage	**3.160**	**.000**	0.205	.095	**1.33×10^{8}**	**.000**	**2.879**	**.000**	**1.204**	**.049**
Centrality Divergence	0.343	.184	**3.679**	**.023**	1.534	.665	**23.400**	**.000**	**7.620**	**.000**
Number of References	**1.013**	**.000**	**1.013**	**.000**	**1.013**	**.000**	**1.012**	**.000**	**1.012**	**.000**
Number of Pages	**0.970**	**.000**	**0.971**	**.000**	**0.971**	**.000**	**0.973**	**.000**	**0.972**	**.000**
Dispersion Parameter (θ)	0.5284		0.5258		0.5150		0.5282		0.5375	
2 x log-likelihood	-31,771		-28,331		-29,491		-29,506		-29,613	
Akaike's Information Criterion (AIC)	31,787		28,347		29,508		29,522		29,629	

In theory, Zero-Inflated Negative Binomial Regression Models (ZINBs) may estimate the effects of variables more accurately, especially in situations where an excessive number of zeros is expected. It is widely known that a considerable number of scientific publications would never be cited. Table 2 summarizes a ZINB model derived from networks of co-cited references (Model 1) and a corresponding NB model (Model 2). The Vuong test indicated that the ZINB model is superior to the NB model at the p-value of 0.0033.

The IRRs of *References* and *Pages* remain identical in both models. The IRRs of Coauthors, MCR, and CL are slightly lower in the ZINB model than that in the NB model. Two variables have statistically significant effects in the zero-inflation model: *MCR* and *References*. In summary, for this particular dataset, *Cluster Linkage* (CL) is a more prominent predictor of global citations than all other more commonly studied factors. Considering the fact that CL is associated with a clearly defined boundary spanning mechanism, this finding provides empirical evidence that such mechanisms are likely to reveal some fundamental insights into the creation of scientific knowledge.

Table 2. ZINB and NB models of global citation counts of 3,515 citing articles on Complex Network Analysis (1996-2004) with reference to networks of co-cited references using $\log_2$(Year of Publication) to account for the exposure time. Coefficients are IRRs. Weighted Cluster Linkage is the strongest predictor of citation counts, followed by the number of co-authors, and the Modularity Change Rate. In this case, with a lower AIC and a statistically significant Vuong test, the ZINB model is superior.

Global Cites	**ZINB**				**NB**	
	Count Model Negbin with log link		Zero-Inflation Model Binomial with logit link			
Coauthors	**1.293**	**.000**	0.062	.077	**1.306**	**.000**
Modularity Change Rate	**1.080**	**.014**	**0.012**	**.044**	**1.083**	**.025**
Weighted Cluster Linkage	**3.103**	**.000**	1.304	.906	**3.160**	**.000**

Centrality Divergence	0.391	.237	385363.6	.102	0.343	.184
Number of References	**1.013**	**.000**	**0.489**	**.027**	**1.013**	**.000**
Number of Pages	**0.970**	**.000**	1.133	.120	**0.970**	**.000**
Dispersion Parameter (θ)	0.536				0.528	
AIC	31,768				31,787	
Vuong Test (ZINB >NB)	-2.7186, p-value=**0.0033**					

Mass Extinction (1991-2010)

The Vuong test was not able to reject the hypothesis that the ZINB model and the NB model are indistinguishable at the p-value of 0.0928. In particular, none of the variables have statistically significant effects on predicting the excessive number of zeros (Table 3).

The ZINB/NB models reveal a similar pattern as we have seen in the Complex Network Analysis dataset. *Cluster Linkage* is again the most prominent predictor of global citation counts, with the strongest IRR coefficient of 3.975 in the ZINB model. The IRRs of *Coauthors* and *References* are slightly greater than 1.0. The IRRs of MCR and Pages are less than 1.0, indicating an expected decrease of citations as the two variables increase independently.

Table 3. ZINB and NB models of global citation counts of 1,745 articles on Mass Extinctions (1991-2010) with reference to networks of co-cited references using $\log_2$(Year of Publication) to account for the exposure time. Coefficients are IRRs. Cluster Linkage is the strongest predictor of citation counts, followed by the number of coauthors, and the number of references. In this case, the ZINB and NB models are indistinguishable statistically.

Global Cites	**ZINB**				**NB**	
	Count Model Negbin with log link		Zero-Inflation Model Binomial with logit link			
Coauthors	**1.160**	**.001**	0.191	.380	**1.162**	**.0006**
Modularity Change Rate	**0.713**	**.021**	9.95×10^{-9}	.269	**0.736**	**.0374**
Weighted Cluster Linkage	**3.975**	**.000**	2.63×10^{5}	.461	**3.805**	**.0000**
Centrality Divergence	9.837	.334	3.65×10^{-20}	.877	8.619	.3607
Number of References	**1.004**	**.000**	1.025	.283	**1.004**	**.0000**
Number of Pages	**0.987**	**.002**	0.657	.116	**0.988**	**.0048**
Dispersion Parameter (θ)	0.489				0.484	
AIC	13,368				13,366	
Vuong Test (ZINB >NB)	-1.3234, p-value=0.0928					

Terrorism (1996-2005)

Unlike the previous cases in which one-year time slices were used, we used time slices with longer exposure time for the Terrorism (1996-2005) dataset, including two- and three-year time slices. A three-year moving window was used with three-year time slices.

The ZINB and NB models revealed similar patterns to other cases we have seen so far, except the strong IRR of *Centrality Divergence*, which was absent in the earlier cases. The effects of *Coauthors*, *CL*, *References*, and *Pages* are within the similar range as shown in earlier cases. Table 4 includes models for both two-year and three-year time slices. The 3-year models have a better relative goodness of fit than the 2-year models according to the AIC scores. The IRR of *Centrality Divergence* is 48.061 in three-year slice models and 62.637 in two-year slice models. We inspected articles with high *Centrality Divergence* in order to find an explanation of the strong *Centrality Divergence*.

Table 4. ZINB and NB models of global citation counts of 3,476 articles on Terrorism (1996-2005) with reference to networks of co-cited references using $\log_2$(Year of Publication) to account for the exposure time. The length of each time slice and the width of the moving window are both three years. Coefficients are IRRs. The strongest predictor is Centrality Divergence, followed by Weighted Cluster Linkage, and the number of coauthors and the number of references. With a lower AIC and a statistically significant Vuong test, the ZINB model is a better model than the NB model.

Slice Length=3 Years						
Global Cites	**ZINB**				**NB**	
	Count Model Negbin with log link		Zero-Inflation Model Binomial with logit link			
Coauthors	**1.808**	**.0000**	0.000	.9524	**1.960**	**.0000**
Modularity Change Rate	0.578	.0973	1.563	.8713	0.581	.0603
Weighted Cluster Linkage	**3.306**	**.0001**	0.000	.7251	**3.300**	**.0002**
Centrality Divergence	**48.061**	**.0012**	0.121	.9368	**64.120**	**.0000**
Number of References	**1.018**	**.0000**	0.980	.0702	**1.019**	**.0000**
Number of Pages	**0.985**	**.0000**	1.002	.8916	**0.986**	**.0048**
Dispersion Parameter (θ)	0.520				0.464	
AIC	20,591				20,619	
Vuong Test (ZINB >NB)	-2.9912, p-value=**0.0014**					
Slice Length=2 Years						
Global Cites	**ZINB**				**NB**	
	Count Model Negbin with log link		Zero-Inflation Model Binomial with logit link			
Coauthors	**1.853**	**.0000**	**0.067**	**.0119**	**2.005**	**.0000**
Modularity Change Rate	**0.586**	**.0000**	0.394	.3708	0.596	.0000
Weighted Cluster Linkage	**3.180**	**.0000**	**28.363**	**.0435**	**2.674**	**.0000**
Centrality Divergence	**62.637**	**.0319**	0.000	.2143	**108.959**	**.0046**
Number of References	**1.018**	**.0000**	0.984	.1296	**1.019**	**.0000**
Number of Pages	**0.986**	**.0000**	1.004	.6902	**0.986**	**.0000**
Dispersion Parameter (θ)	0.518				0.464	
AIC	20,668				20,695	
Vuong Test (ZINB >NB)	-2.8463, p-value=**0.0022**					

Figure 6 illustrates connections made by the top five articles with the strongest centrality divergence scores based on two-year time slice models. As show in the figure, much of the boundary-spanning activities introduced by this group of articles involved three major clusters, namely, Cluster #12 biological terrorism, Cluster #13 ocular injury, and Cluster #5 September 11.

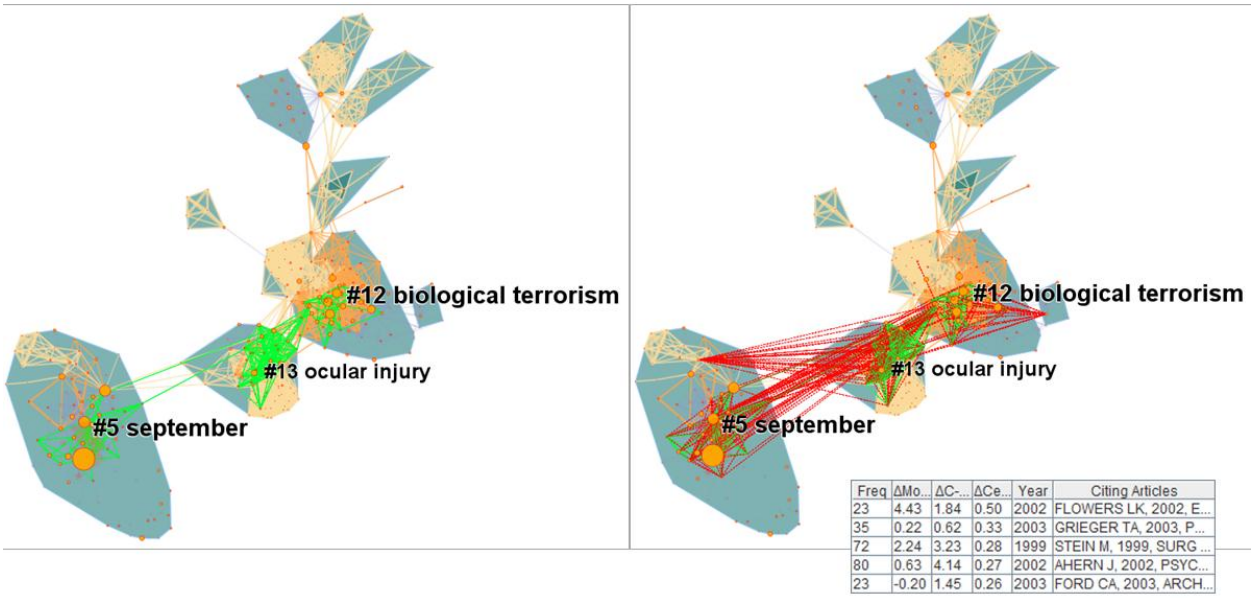

Figure 6. Co-citations made by the five articles with the strongest centrality divergence scores in Terrorism (1996-2005) that reinforced existing patterns (Left) introduced novel connections (Right). Structural variations were computed based on two-year time slices.

The five top-ranked high centrality divergence articles are listed in Table 5, including their DOIs. One can tell from their titles which topic areas they were connecting, for example, connections between terrorism and medical services (papers #1 and #3) and connections between post-traumatic stress disorder and September 11 terrorist attacks (papers #2, #4, and #5). In terms of the clusters shown in Figure 6, the topic of bioterrorism is clearly associated with Cluster #12, the topic of September 11 terrorist attacks is the focus of Cluster #5, and the topic of clinical and medical services is closely related to Cluster #13.

The current analysis suggests that the strong IRR of *Centrality Divergence* may be due to the wide-ranging interdisciplinary structure of the subject matter. If this is indeed the case, then a strong *Centrality Divergence* can be a valuable early sign of transformative research at interdisciplinary levels. Further research is needed to gain additional insights into implications of these new structural variation metrics.

Table 5. The top 5 articles in Terrorism (1996-2005) with the strongest centrality divergence scores. Co-citation connections made by these papers are shown in Figure 6.

#	Cites	C_{KL}	Author	Year	Title. *Source*. DOI
1	23	0.500	FLOWERS LK	2002	Bioterrorism preparedness II: the community and emergency medical services systems. *EMERG MED CLIN N AM*. 10.1016/S0733-8627(01)00009-8
2	35	0.327	GRIEGER TA	2003	Posttraumatic stress disorder alcohol use and perceived safety after the terrorist attack on the Pentagon. *PSYCHIAT SERV*. 10.1176/appi.ps.54.10.1380
3	72	0.285	STEIN M	1999	Medical consequences of terrorism - The conventional weapon threat. *SURG CLIN N AM*. 10.1016/S0039-6109(05)70091-8
4	80	0.271	AHERN J	2002	Television images and psychological symptoms after the September 11 terrorist attacks. *PSYCHIATRY*. 10.1521/psyc.65.4.289.20240

5	23	0.264	FORD CA	2003	Reactions of young adults to September 11 2001. *ARCH PEDIAT ADOL MED*. 10.1001/archpedi.157.6.572

CiteSpace Expanded (2001-2010)

This dataset was formed with a seed article (Chen, 2006) by including papers that have at least one cited reference in common with the seed article during the period of 2001 and 2010. Estimates made by ZINB and NB models are listed in Table 6. The Vuong test was not able to reject the null hypothesis that the ZINB and NB models are indistinguishable. A zero-inflated negative binomial regression model shows a statistically significant main effect of the cluster linkage variation with an IRR of 3.230. *References* and *Pages* have similar values as in other cases. The NB model reveals a strong effect of *Centrality Divergence* (an IRR of 3.453).

Table 6. ZINB and NB models of global citation counts of 3,260 articles published between 2001 and 2010 in an area seeded by our 2006 JASIST article with reference to networks of co-cited references using $\log_2$(Year of Publication) to offset the exposure time. Weighted Cluster Linkage is the strongest predictor of citation counts, followed by the number of references. In this case, the ZINB and NB models are not indistinguishable statistically.

Global Cites	**ZINB**				**NB**	
	Count Model Negbin with log link		Zero-Inflation Model Binomial with logit link			
Coauthors	1.070	.0826	4.125	.3379	**1.084**	**.0225**
Modularity Change Rate	1.025	.5482	**1.715**	**.0198**	1.013	.7135
Weighted Cluster Linkage	**3.230**	**.0046**	0.001	.0834	**3.589**	**.0000**
Centrality Divergence	3.203	.0870	0.000	.1590	**3.453**	**.0272**
Number of References	**1.010**	**.0000**	**0.958**	**.0124**	**1.010**	**.0000**
Number of Pages	**0.985**	**.0000**	1.097	.0844	**0.984**	**.0048**
Dispersion Parameter (θ)	0.423				0.413	
AIC	19,797				19,786	
Vuong Test (ZINB >NB)	-0.4694, p-value=0.3194					

The seed article describes CiteSpace, a visual analytic tool for identifying emerging trends and changes in scientific literature. It has been cited 100 times in the Web of Science and has the highest CL score (12.78) and the highest Centrality Divergence score (1.24).

Discussions

ZINB models are superior to their NB counterparts in three cases we studied, but indistinguishable in the other two cases (See Table 7). The AIC values of these models increase as the number of citing articles per time slices increases, but there is no apparent trend in connection to whether the datasets are seeded or not. Mass Extinctions and CiteSpace Expanded have the lowest AIC values, whereas Complex Networks has the highest AIC value. In the Terrorism case, the use of three-year time slices improved the models over the two-year configurations. These results seem to suggest that ZINB models are particularly recommended as the size of a dataset increases. As the number of articles increases, more and more zero-citation articles need to be account for.

Table 7. AICs of ZINB and NB models. ZINBs are distinguishable from NB models in three of the five cases. AICs increase as the number of papers per time slice increases.

Cases	Seeded	From	Years	N	N/Year	AIC(ZINB)	AIC(NB)	ZINB>NB (p)
Mass Extinctions	No	1991-2010	20	1745	87.25	13368	**13366**	0.0928
CiteSpace Expanded	Yes	2001-2010	10	3260	163.00	19797	**19786**	0.3194
Terrorism (3-year slices)	No	1996-2005	10	3476	173.80	**20591**	20619	**0.0014**
Terrorism (2-year slices)	No	1996-2005	10	3476	173.80	**20668**	20695	**0.0022**
Complex Networks	Yes	1996-2004	9	3515	175.75	**31768**	31787	**0.0033**

The Incidence Rate Ratio (IRR) of a variable has an intuitive interpretation. If it is greater than 1.0 and all other variables are held constant, then the citation count is expected to increase along with a unit increase of the variable. In contrast, if it is less than 1.0, then the citation count is expected to decrease as the variable increases. *Cluster Linkage*, *Coauthors*, and *References* are persistently found to predict an increase of citation counts. The strongest predictor, *Cluster Linkage*, has an average IRR of 3.431, which is more than twice of the second strongest predictor, *Coauthors* (the average IRR of 1.440), and more than three times of the third strongest predictor, *References* (the average IRR of 1.013). The length of an article in terms of the number of pages has a minor but consistent negative impact with an average IRR of 0.982 across all five cases.

In contrast, the IRR estimates of *Centrality Divergence* and *MCR* did not give clear signals. Although the largest average IRR (24.876) is found with *Centrality Divergence* and in three of the five cases the effect was positive and statistically significant, it has a negative impact on citations in the Complex Network case. It is possible that *Centrality Divergence* is sensitive to the structure of the underlying network as well as to the structural change of the network. Given its strong IRRs in its magnitude and the possible sensitivity, the behavior of the metric should be investigated further in order to explain the discrepancies across different cases.

MCR is an indirect measure of the impact of cross-cluster links. It has an average IRR of 0.794, but the estimated IRRs are inconsistent across the five cases in terms of statistical significance and the direction of citation change. The discrepancies may reflect the connectivity of individual nodes involved. For example, adding a new link to a node with either a large degree or a small degree is likely to affect the MCR much more than linking two nodes of similar connectivity.

Table 8. Incidence Rate Ratios (IRRs) of predictors of global citation counts. IRRs in bold font are statistically significant at the p-level of 0.05. IRRs of Coauthors, Cluster Linkage, References, and Pages are consistent across the cases, whereas IRRs of MCR and Centrality Divergence are mixed.

Cases	Coauthors	MCR	Cluster Linkage	Centrality Divergence	References	Length in Pages
Mass Extinctions	**1.160**	**0.713**	**3.975**	9.837	**1.004**	**0.987**
CiteSpace Expanded	**1.084**	1.013	**3.589**	**3.453**	**1.010**	**0.984**
Terrorism (3-year slices)	**1.808**	0.578	**3.306**	**48.061**	**1.018**	**0.985**
Terrorism (2-year slices)	**1.853**	**0.586**	**3.180**	**62.637**	**1.018**	**0.986**

Complex Networks	**1.293**	**1.080**	**3.103**	0.391	**1.013**	**0.970**
Mean	**1.440**	**0.794**	**3.431**	24.876	**1.013**	**0.982**

Cluster Linkage is a more direct measure of intellectual potential than *Coauthors* and *References*. *Cluster Linkage* has a simple and clear theoretical interpretation supported by the underlying boundary-spanning mechanism. The strong IRRs of *Cluster Linkage* suggest that the structural variation metric is more efficient to predict the growth of citation counts than indicators such as *Coauthors* and *References*.

These findings are very encouraging because of their theoretical and practical implications. These findings are valuable in improving our understanding of how transformative ideas can be made and recognized. The premise of the structural variation approach is that transformative ideas are expected to introduce significant and computationally detectable structural changes. The value of such approaches has been identified by many researchers (Boyack, et al., 2005; Leydesdorff, 2001; Shibata, et al., 2007; Upham, et al., 2010; van Dalen & Kenkens, 2005). The boundary-spanning mechanism provides one possible explanation of what factors attract citations. It is conceivable that an idea has the potential to introduce novel structural changes, but fails to materialize it due to many reasons. In the present study, we use the co-citation network immediately prior to the publication of a new paper as the reference point. There are several other options at various levels of granularity and scale. For instance, in the Complex Network Analysis case, structural variation models based on networks of co-occurring keywords have the best AIC score with an IRR of *Centrality Divergence* of 3.679. Models based on networks of co-cited authors led to a strong *Centrality Divergence* IRR of 23.400. We do not have the room to investigate the implications of these results in the present article, but we believe these results indicate a wide range of potentially significant directions of research that is far beyond the scope of a single study.

Conclusions

We have found statistical evidence of the boundary-spanning mechanism. An article that introduces novel connections between clusters of co-cited references is likely to become highly cited subsequently. In addition, we have found that the IRRs of *Cluster Linkage* are more than twice as much as the IRRs of *Coauthors* and *References*. This finding provides a more fundamental explanation of why the number of references cited by an article appears to be a good predictor of its future citations as found in many previous studies. As a result, the structural variation paradigm clarifies why a number of extrinsic features appear to be associated with high citations.

A distinct characteristic of the structural variation approach is the focus on the potential connection between the degree of structural variation introduced by an article and its future impact. The analytic and modeling procedure demonstrated in this article is expected to serve as an exemplar for subsequent studies along this line of research. More importantly, the focus on the underlying mechanisms of scientific activity is expected to provide additional insights and practical guidance for scientists, sociologists, historians, and philosophers of scientific knowledge.

There are many new challenges and opportunities ahead. For example, how common is the boundary-spanning mechanism in scientific discoveries overall? What are the other major mechanisms and how do they interact with the boundary-spanning mechanism? There are other potentially valuable techniques that

we have not utilized in the present study, including topic modeling, citation context analysis, survival analysis and burst detection. In short, a lot of work is to be done and this is an encouraging start.

We conclude that structural variation is an essential aspect of the development of scientific knowledge and it has the potential to reveal the underlying mechanisms of the growth of scientific knowledge. The focus on the underlying mechanisms of knowledge creation is the key to the predictive potential of the structural variation approach. The theory-driven explanatory and computational approach sets an extensible framework for detecting and tracking potentially creative ideas and gaining insights into challenges and opportunities in light of the collective wisdom.

Acknowledgements

The author would like to thank anonymous reviewers for their valuable comments.

REFERENCES

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation, 12*(3), 159-170.

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509-512.

Bornmann, L., & Daniel, H.-D. (2006). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45-80.

Boyack, K. W., Klavans, R., Ingwersen, P., & Larsen, B. (2005). *Predicting the importance of current papers*. Paper presented at the Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics. Retrieved from https://cfwebprod.sandia.gov/cfdocs/CCIM/docs/kwb_rk_ISSI05b.pdf

Brody, T., & Harnad, S. (2005). Earlier web usage statistics as predictors of later citation impact. from http://arxiv.org/ftp/cs/papers/0503/0503020.pdf

Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology, 110*(2), 349-399.

Buter, R., Noyons, E., & Van Raan, A. (2011). Searching for converging research using field to field citations. *Scientometrics, 86*(2), 325-338.

Chen, C. (2003). *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. London: Springer-Verlag.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359-377.

Chen, C. (2011). *Turning Points: The Nature of Creativity*: Springer.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics, 3*(3), 191-209.

Chen, C., Cribbin, T., Macredie, R., & Morar, S. (2002). Visualizing and tracking the growth of competing paradigms: Two case studies. *Journal of the American Society for Information Science and Technology, 53*(8), 678-689.

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The Structure and Dynamics of Co-Citation Clusters: A Multiple-Perspective Co-Citation Analysis. *Journal of the American Society for Information Science and Technology, 61*(7), 1386-1409.

Chen, C., Lin, X., & Zhu, W. (2006). *Trailblazing through a knowledge space of science: Forward citation expansion in CiteSeer*. Paper presented at the Proceedings of the 69th Annual Meeting of the American Society for Information Science and Technology (ASIS&T 2006). Retrieved from http://eprints.rclis.org/archive/00008019/

Chubin, D. E. (1994). Grants peer-review in theory and practice. *Evaluation Review, 18*(1), 20-30.

Chubin, D. E., & Hackett, E. J. (1990). *Paperless science: peer review and U.S. science policy*.

Cuhls, K. (2001). Foresight with Delphi surveys in Japan. [Article]. *Technology Analysis & Strategic Management, 13*(4), 555-569.

Dewett, T., & Denisi, A. S. (2004). Exploring scholarly reputation: It's more than just productivity. [Article]. *Scientometrics, 60*(2), 249-272.

Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. *Cognitive Science, 22*(2), 133-187.

Fleming, L., & Bromiley, P. (2000). *A variable risk propensity model of technological risk taking*. Paper presented at the Applied Statistics Workshop. Retrieved from http://courses.gov.harvard.edu/gov3009/fall00/fleming.pdf

Galea, S., Ahern, J., Resnick, H., Kilpatrick, D., Bucuvalas, M., Gold, J., et al. (2002). Psychological sequelae of the September 11 terrorist attacks in New York City. *New England Journal of Medicine, 346*(13), 982–987.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science, 122*(108-111).

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., et al. (2008). What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics, 76*(1), 169-185.

Häyrynen, M. (2007). *Breakthrough research: funding for high-risk research at the Academy of Finland*. Helsinki: The Academy of Finland.

Hettich, S., & Pazzani, M. J. (2006). *Mining for proposal reviewers: Lessons learned at the National Science Foundation*. Paper presented at the KDD'06.

Hilbe, J. M. (2011). *Negative Binomial Regression* (2nd ed.): Cambridge University Press.

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences, 104*(49), 19193-19198.

Hsieh, C. (2011). Explicitly searching for useful inventions: dynamic relatedness and the costs of connecting versus synthesizing. *Scientometrics, 86*(2), 381-404.

Kostoff, R. (2007). The difference between highly and poorly cited medical articles in the journal <i>Lancet</i&gt. *Scientometrics, 72*, 513-520.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., et al. (2005). The effect of use and access on citations. *Information Processing & Management, 41*(6), 1395-1402.

Lahiri, M., Maiya, A. S., Sulo, R., Habiba, & Wolf, T. Y. B. (2008). *The impact of structural changes on predictions of diffusion in networks*. Paper presented at the The 2008 IEEE International Conference on Data Mining Workshops (ICDMW '08). Retrieved from http://compbio.cs.uic.edu/~mayank/papers/LahiriMaiyaSuloHabibaBergerWolf_ImpactOfStructuralChanges08.pdf

Lambert, D. (1992). Zero-infated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34*, 1-14.

Levitt, J., & Thelwall, M. (2008). Patterns of annual citation of highly cited articles and the prediction of their citation ranking: A comparison across subjects. *Scientometrics, 77*(1), 41-60.

Leydesdorff, L. (2001). *The Challenge of Scientometrics: The Development, Measurement, and Self-Organization of Scientific Communications*: Universal-Publishers.

Lipinski, C., & Hopkins, A. (2004). Navigating chemical space for biology and medicine. [Article]. *Nature, 432*(7019), 855-861.

Luxburg, U. v. (2006). A tutorial on spectral clustering. from http://www.kyb.mpg.de/publications/attachments/Luxburg06_TR_%5B0%5D.pdf

Martin, B. R. (2010). The origins of the concept of 'foresight' in science and technology: An insider's perspective. *Technological Forecasting and Social Change, 77*(9), 1438-1447.

Merton, R. K. (1968). The Mathew Effect in science. *Science, 159*(3810), 56-63.

Miles, I. (2010). The development of technology foresight: A review. *Technological Forecasting and Social Change, 77*(9), 1448-1456.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America, 103*(23), 8577-8582.

Persson, O. (2010). Are highly cited papers more international? *Scientometrics, 83*(2), 397-401.

Price, D. D. (1965). Networks of scientific papers. *Science, 149*, 510-515.

Shibata, N., Kajikawa, Y., & Matsushima, K. (2007). Topological analysis of citation networks to discover the future core articles. *Journal of the American Society for Information Science and Technology, 58*(6), 872-882.

Skilton, P. (2009). Does the human capital of teams of natural science authors predict citation frequency? *Scientometrics, 78*(3), 525-542.

Swanson, D. R. (1986a). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*(30), 7-18.

Swanson, D. R. (1986b). Undiscovered public knowledge. *Library Quarterly, 56*(2), 103-118.

Takeda, Y., & Kajikawa, Y. (2010). Tracking modularity in citation networks. *Scientometrics, 83*(3), 783.

Tichy, G. (2004). The over-optimism among experts in assessment and foresight. [Article]. *Technological Forecasting and Social Change, 71*(4), 341-363.

Tijssen, R. J. W., Visser, M. S., & van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? [Article]. *Scientometrics, 54*(3), 381-397.

Upham, S. P., Rosenkopf, L., & Ungar, L. H. (2010). Positioning knowledge: schools of thought and new knowledge creation. *Scientometrics, 83*, 555-581.

van Dalen, H. P., & Kenkens, K. (2005). Signals in science: On the importance of signaling in gaining attention in science. *Scientometrics, 64*(2), 209-233.

Walters, G. D. (2006). Predicting subsequent citations to articles published in twelve crime-psychology journals: Author impact versus journal impact. *Scientometrics, 69*(3), 499-510.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440-442.

Weeber, M. (2003). Advances in literature-based discovery. *Journal of the American Society for Information Science and Technology, 54*(10), 913-925.

Zeileis, A., Kleiber, C., & Jackman, S. (2011). Regression models for count data in R. from http://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf