A Visual Analytic Study of Retracted Articles in Scientific Literature

Chaomei Chen[1], Zhigang Hu[2], Jared Milbank[3], Timothy Schultz[1]
[1]College of Information Science and Technology, Drexel University
3141 Chestnut Street, Philadelphia, PA 19085, USA
[2]WISELab, Dalian University, Dalian, China
[3]Pfizer Worldwide Research & Development, Groton Laboratories
Eastern Point Rd, Groton, CT 06340, USA
Email: chaomei.chen@drexel.edu

Abstract
Retracting published scientific articles is increasingly common. Retraction is a self-correction mechanism of the scientific community to maintain and safeguard the integrity of scientific literature. However, a retracted article may pose a profound and long-lasting threat to the credibility of the literature. New articles may unknowingly build their work on false claims made in retracted articles. Such dependencies on retracted articles may become implicit and indirect. Consequently, it becomes increasingly challenging to detect implicit and indirect threats. In this article, our aim is to raise the awareness of the potential threats of retracted articles even after their retraction and demonstrate a visual analytic study of retracted articles with reference to the rest of the literature and how their citations are influenced by their retraction. The context of highly cited retracted articles is visualized in terms of a co-citation network as well as the distribution of articles that have high-order citation dependencies on retracted articles. Survival analyses of time to retraction and post-retraction citation are included. Sentences that explicitly cited retracted articles are extracted from full text articles. Transitions of topics over time are depicted in topic-flow visualizations. We recommend that new visual analytic and science mapping tools should take retracted articles into account and facilitate tasks specifically related to the detection and monitoring of retracted articles.

Keywords: Visual analytics, study of retracted articles, citation analysis, network visualization

Introduction

The reproducibility of findings reported in scientific publications is a major hallmark of the validity of science. If fellow scientists follow the same procedure described in a scientific publication, they would expect to be able to reproduce the findings in the original publication to a great extent. However, if numerous attempts by different scientists could not reproduce the original findings, then scientists may start to question the validity of the original publication. The retraction of a scientific article is a formal action that is taken to purge the article from the scientific literature on the ground that the article in question is not trustworthy and therefore no longer qualified to be part of the intellectual basis of scientific knowledge.

Retraction is a self-correct mechanism of the scientific community. Scientific articles can be retracted for a variety of reasons, ranging from self-plagiarism, editorial errors, to scientific misconduct, which may include fabrication and falsification of data and results. The consequences of these diverse types of mistakes differ. Some are easier to detect than others. For example, clinical studies contaminated by fabrications of data or results may directly risk the safety of patients, whereas publishing a set of valid results simultaneously in multiple journals is not ethical but nonetheless less likely to harm patients directly. Some retracted articles may remain to be controversial even after their retraction. For example, Lancet partially retracted a 1998 paper (Wakefield et al., 1998) that suggested a possible link between a

combination of vaccines against measles, mumps, and rubella and autism. The ultimate full retraction of the Lancet article didn't come until 2010. On the other hand, the influence of other retracted articles may come to an end more abruptly after their retraction, for example, the fabricated stem cell clone by Woo-Suk Hwang (Kakuk, 2009). Monitoring retractions of scientific article is an important part of the current practice with notable examples such as Retraction Watch[1].

Monitoring and tracking retracted articles has become increasingly challenging. The rate of retraction from the scientific literature has been steadily increasing. For example, retractions in MEDLINE have increased sharply since 1980. Reasons for retraction include errors or non-reproducible findings (40%), research misconduct (28%), redundant publication (17%) and unstated/unclear (5%) (Wager & Williams, 2011). Figure 1 is a snapshot of the status of PubMed as of 3/29/2012. The total number of annual publications in PubMed increased from slightly more than 543,000 articles in 2001 to more than 984,000 articles in 2011. The increase has been remarkably steady, adding about 45,000 new articles per year. The number of retracted articles in a year refers to the number of articles that are published in that year but subsequently retracted. The rate of retraction is the number of retraction notices issued each year divided by the total number of new publications added to PubMed in the same year. The retraction rate in 2001 was 0.00005. It was doubled three times since then, in 2003, 2006, and 2011, respectively. The retraction rate in 2011 was 0.00046. Figure 1 also shows that the number of retracted articles per year peaked in 2006. The blue line is the retraction rate, which is growing up fast. The red line is the actual number of retracted articles. Although fewer articles have been retracted in recent years than the 2006 peak number, we expect that these numbers will continue to grow because recognizing potential flaws in newly published articles lag behind their initial publications. Later in this article, we will provide estimates of such delays in terms of how long a retraction is most likely to occur and how likely for a retracted article to be cited after its retraction.

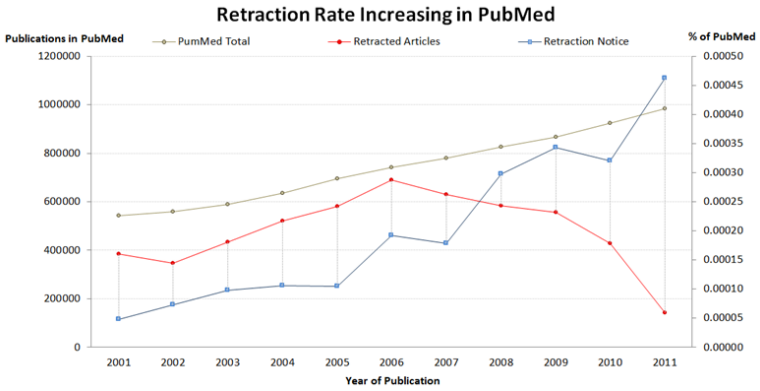


Figure 1. The rate of retraction is increasing in PubMed (As of 3/29/2012).

In light of the urgency and severity of potential consequences, the study of retracted articles is still at a relatively early stage and yet to establish its integral role in serving scientific communities. The awareness of mistakes in scientific studies has been improving (Naik, 2011), especially due to the publicity of high-profile retraction and fraudulent cases (Kakuk, 2009; Service, 2002). However, many profound issues need to be addressed in a broader context over a longer period of time than what is available in the contemporary literature. In this article, our primary goal is twofold: 1) to identify the extent to which retracted articles are interwoven with the rest of the scientific literature in terms of how they are tightly embedded in co-citation networks, and 2) to demonstrate the potential of a visual analytics approach that can be used by a broad range of researchers and analysts to examine and monitor not only retracted articles per se but also articles that might be at risk of contamination. We also aim to demonstrate how

[1] http://retractionwatch.wordpress.com/

sense making tasks for situation awareness may be supported at multiple levels of granularity, ranging from interrelated topics at a macroscopic level down to how a retracted article is cited before and after its retraction at a microscopic level of sentences found in full text articles.

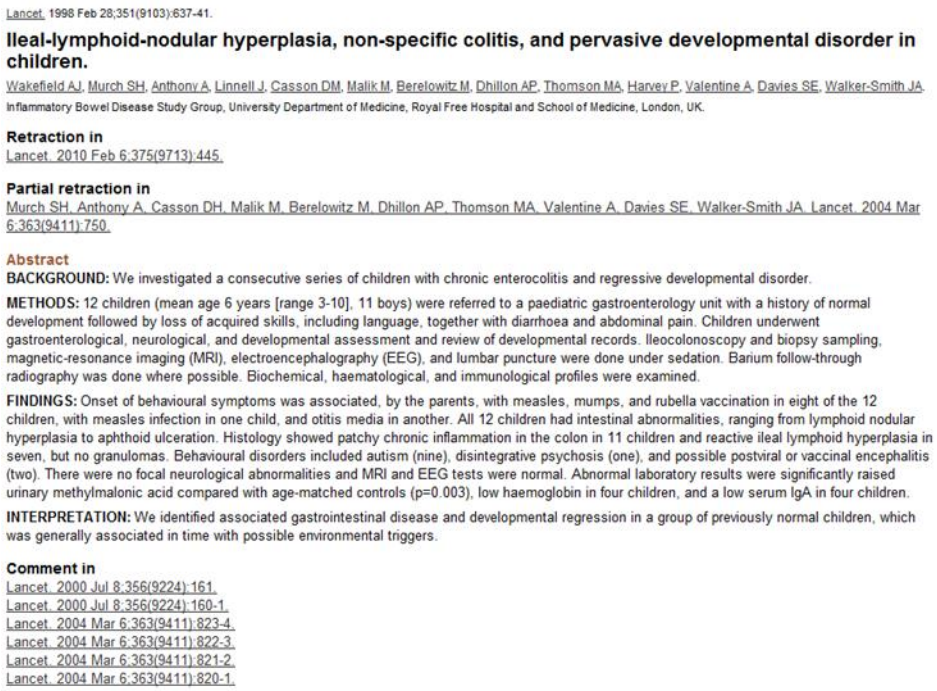
Related Work

Retraction is considered as the most serious sanction that can be applied to a scientific publication (Steen, 2011). Studies of retracted articles typically address a series of common questions concerning retraction. For example, what are the most common reasons for retracting an article? How long on average does a retraction take place since the initial publication of the article in question? How often is a retracted article cited after its retraction? Existing studies have not particularly focused on a higher-order of impact of a retracted article in a broader context in terms of structural and temporal patterns and properties. How tightly is a retracted article interwoven in a network of other articles? How often is attention paid to the need of re-examining the validity of these other articles? What should be done to articles that build on an ultimately retracted article? What analytic tools are needed to support tasks for monitoring and verifying the impact of a retracted article? In the following section, we will outline findings of existing studies in the current literature and highlight questions that still need to be addressed.

Finding Retracted Articles

How do we find out whether an article has been retracted? The retraction of an article is officially announced in a retraction notice. We describe how retracted articles can be identified in PubMed, the Web of Science, and Google Scholar.

PubMed is the largest publically available resource of the scientific literature with the most extensive coverage of scientific publications in medicine and related disciplines. The Publication Type [pt] of the record of a retraction notice is "Retraction of Publication." The Publication Type of the record of the original article is updated to "Retracted Publication." PubMed provides a list of special queries, including one for "retracted publication."[2] Figure 2 illustrates the history of the retraction of the Wakefield paper we mentioned earlier, which was partially retracted in 2004 and fully retracted in 2010.



Lancet. 1998 Feb 28;351(9103):637-41.

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children.

Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, Berelowitz M, Dhillon AP, Thomson MA, Harvey P, Valentine A, Davies SE, Walker-Smith JA.

Inflammatory Bowel Disease Study Group, University Department of Medicine, Royal Free Hospital and School of Medicine, London, UK.

Retraction in

Lancet. 2010 Feb 6;375(9713):445.

Partial retraction in

Murch SH, Anthony A, Casson DH, Malik M, Berelowitz M, Dhillon AP, Thomson MA, Valentine A, Davies SE, Walker-Smith JA. Lancet. 2004 Mar 6;363(9411):750.

Abstract

BACKGROUND: We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

METHODS: 12 children (mean age 6 years [range 3-10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

FINDINGS: Onset of behavioural symptoms was associated, by the parents, with measles, mumps, and rubella vaccination in eight of the 12 children, with measles infection in one child, and otitis media in another. All 12 children had intestinal abnormalities, ranging from lymphoid nodular hyperplasia to aphthoid ulceration. Histology showed patchy chronic inflammation in the colon in 11 children and reactive ileal lymphoid hyperplasia in seven, but no granulomas. Behavioural disorders included autism (nine), disintegrative psychosis (one), and possible postviral or vaccinal encephalitis (two). There were no focal neurological abnormalities and MRI and EEG tests were normal. Abnormal laboratory results were significantly raised urinary methylmalonic acid compared with age-matched controls ($p=0.003$), low haemoglobin in four children, and a low serum IgA in four children.

INTERPRETATION: We identified associated gastrointestinal disease and developmental regression in a group of previously normal children, which was generally associated in time with possible environmental triggers.

Comment in

Lancet. 2000 Jul 8;356(9224):161.
Lancet. 2000 Jul 8;356(9224):160-1.
Lancet. 2004 Mar 6;363(9411):823-4.
Lancet. 2004 Mar 6;363(9411):822-3.
Lancet. 2004 Mar 6;363(9411):821-2.
Lancet. 2004 Mar 6;363(9411):820-1.

[2] http://www.ncbi.nlm.nih.gov/PubMed?term=retracted+publication+[pt]

Figure 2. A retracted article with a partial retraction and a full retraction. http://www.ncbi.nlm.nih.gov/pubmed/9500320

Similarly to the Publication Type in PubMed, the Web of Science defines the Document Type a bibliographic record in terms of Article, Review, Correction, and a few other types. The type Correction[3] is used for retractions as well as other types of corrections such as additions and errata (See Figure 3). The title of a retraction notice identifies the article to be retracted by its title and a phrase "(Retraction of)." The title of the original article in the Web of Science is modified accordingly to indicate the fact that the article has been retracted. For example, the Wakefield paper is shown with a phrase "(Retracted article. See vol 375, pg 445, 2010)" (see Figure 4).

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children (Retraction of vol 351, pg 637, 1998)

Author(s): Wakefield, AJ (Wakefield, A. J.); Murch, SH (Murch, S. H.); Anthony, A (Anthony, A.)
Source: LANCET **Volume:** 375 **Issue:** 9713 **Pages:** 445-445 **DOI:** 10.1016/S0140-6736(10)60175-4 **Published:** FEB 6 2010
Times Cited: 7 (from Web of Science)
Cited References: 2 [view related records] Citation Map
Accession Number: WOS:000274757900010
Document Type: Correction
Language: English
Publisher: ELSEVIER SCIENCE INC, 360 PARK AVE SOUTH, NEW YORK, NY 10010-1710 USA
Web of Science Category: Medicine, General & Internal
Subject Category: General & Internal Medicine
IDS Number: 558PV
ISSN: 0140-6736

Figure 3. The retraction notice of the Wakefield paper in the Web of Science.

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children (Retracted article. See vol 375, pg 445, 2010)

Author(s): Wakefield, AJ (Wakefield, AJ); Murch, SH (Murch, SH); Anthony, A (Anthony, A); Linnell, J (Linnell, J); Casson, DM (Casson, DM); Malik, M (Malik, M); Berelowitz, M (Berelowitz, M); Dhillon, AP (Dhillon, AP); Thomson, MA (Thomson, MA); Harvey, P (Harvey, P); Valentine, A (Valentine, A); Davies, SE (Davies, SE); Walker-Smith, JA (Walker-Smith, JA)
Source: LANCET **Volume:** 351 **Issue:** 9103 **Pages:** 637-641 **DOI:** 10.1016/S0140-6736(97)11096-0 **Published:** FEB 28 1998
Times Cited: 742 (from Web of Science)
Cited References: 26 [view related records] Citation Map
Abstract: Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3-10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Findings Onset of behavioural symptoms was associated, by the parents, with measles, mumps, and rubella vaccination in eight of the 12 children, with measles infection in one child, and otitis media in another. All 12 children had intestinal abnormalities, ranging from lymphoid nodular hyperplasia to aphthoid ulceration. Histology showed patchy chronic inflammation in the colon in 11 children and reactive ileal lymphoid hyperplasia in seven, but no granulomas. Behavioural disorders included autism (nine), disintegrative psychosis (one), and possible postviral or vaccinal encephalitis (two). There were no focal neurological abnormalities and MRI and EEG tests were normal. Abnormal laboratory results were significantly raised urinary methylmalonic acid compared with age-matched controls (p=0.003), low haemoglobin in four children, and a low serum IgA in four children.

Interpretation We identified associated gastrointestinal disease and developmental regression in a group of previously normal children, which was generally associated in time with possible environmental triggers.
Accession Number: WOS:000072364200010
Document Type: Article
Language: English
KeyWords Plus: CROHNS-DISEASE; MEASLES VACCINATION; AUTISM
Reprint Address: Wakefield, AJ (reprint author), Univ London Royal Free Hosp, Sch Med, Dept Med, Inflammatory Bowel Dis Study Grp, London NW3 2QG, England

Figure 4. The retracted Wakefield paper as shown in the Web of Science.

[3] **Correction:** Correction of errors found in articles that were previously published and which have been made known after that article was published. Includes additions, errata, and retractions. http://images.webofknowledge.com/WOKRS51B6/help/WOS/hs_document_type.html

In Google Scholar, retracted articles are identified with a prefix of "RETRACTED ARTICLE" to their title (see Figure 5). In advanced Scholar search, one can limit the search to all the records with the phrase in the title.

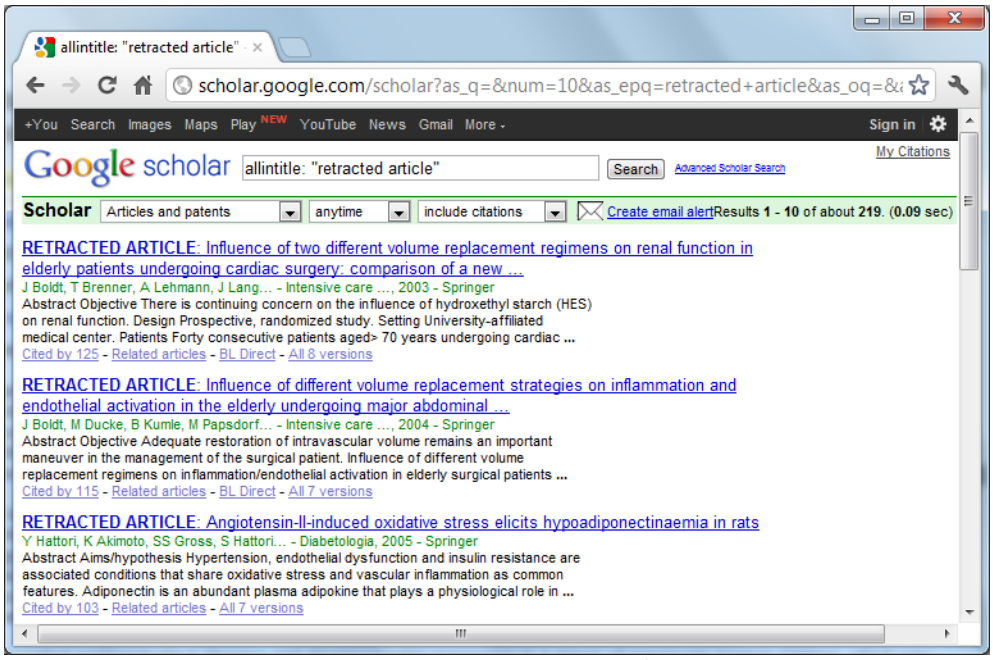


Figure 5. Google Scholar tags retracted articles with a prefix RETRACTED ARTICLE.

Table 1 summarizes the number of retractions found in major sources of scientific publications as of 3/29/2012. The search on PubMed contains all the years available, whereas the search on the Web of Science is limited by the coverage of our institutional subscription (1980 – present).

Table 1. The number of retractions found in major sources of scientific publications (As of 3/29/2012).

Sources	Items	Document Type	Search Criteria
PubMed[4]	2,073	Retracted Article	"Retracted Publication" [pt]
	2,187	Retraction Notice	"Retraction of Publication" [pt]
Web of Science (1980-present)	1,775	Retracted Article	Title contains "(Retracted article."
	1,734	Retraction Notice	Title contains "(Retraction of vol"
Google Scholar	219	Retracted Article	allintitle: "retracted article"
Elsevier Content Syndication (CONSYN)	659	Retracted Article (Full Text)	Title: Retracted Article

Reasons for Retraction and Related Findings

A retraction sends a strong signal to the scientific community that retracted articles are no longer considered trustworthy and they should be effectively purged from the literature. Studies of retraction have typically focused on formally retracted articles. Some have suggested that retraction should be only used to deal with scientific misconduct (Sox & Rennle, 2006). It is believed that many more articles could and should have been retracted (Steen, 2011). The following questions are commonly raised in studies of retraction:

[4] http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=DetailsSearch&Term=%22retracted+publication%22%5Bpublication+type%5D

- *Reasons for retraction* – What are the most common reasons that lead to the retraction of an article? How effective does a retraction serve these reasons? Do scientists simply make mistakes with good faith or some of them intended to cheat with deliberate misconduct?
- *Time to retraction* – How long does it take on average from the publication of a scientific article to its retraction? What are the factors that may influence the time to retraction?
- *Post-retraction citations*– How does the retraction of an article affect citations to the article? What are the reasons for continuously citing a retracted article?
- *Cause of concern* – How was an eventually retracted article noticed in the first place? Are there any early signs that one can watch for and safeguard the integrity of scientific publications? What are the possible ways that a retracted article may damage the scientific literature?

Some of the most representative findings in the literature are summarized in Table 2. The most common causes of initial concern include irreproducibility and an unusually high-level of productivity. For example, Jan Hendrik Schön published a new paper every 8 days during his peak time and fabricated 17 papers in 2 years in both *Science* and in *Nature* (Steen, 2011). Irreproducibility can be caused by a spectrum of more specific reasons, including technical errors and deliberate misconduct. It has been argued that, pragmatically speaking, fabricating data and manipulating results is perceived to be much more harmful than plagiarizing a description or an expression. For example, some researchers suggest that data plagiarism is a more damaging scientific misconduct than text plagiarism (Steen, 2011).

Several studies found that it takes about two years on average to retract a scientific publication. It could take even longer for articles authored by senior researchers. Survival analysis has been used to analyze time to retraction, especially to estimate the probability that an article could survive as a function of time elapsed since its publication (Trikalinos, Evangelou, & Ioannidis, 2008). Based on retractions made in top-cited high-impact journals, it was found that the median survival time of eventually retracted articles was 28 months. In addition, it took much longer to retract articles authored by senior researchers than junior ones. Senior researchers include professors, lab directors, or researchers with more than 5 years of publication records.

Post-retraction citations refer to citations to a retracted article. Some studies started to count citations from the next calendar year of the retraction date (Pfeifer & Snodgrass, 1990), whereas other studies did not start counting citations until 1 year after the retraction (Budd, Sievert, & Schultz, 1998) or 3 years after (Neale, Northrup, Dailey, Marks, & Abrams, 2007). Existing studies show that post-retraction citations do decrease over time, but in some cases post-retraction citations can last as long as 23 years after the retraction.

One way that may differentiate an intended fraudulent behavior from a good faith mistake is to see how often the same researcher has been found with the similar problems. A frequent offender is more likely to have done it deliberately. Studies have indeed found a higher rate of repeat offenders in fraudulent papers than in erroneous papers (Steen, 2011).

Existing studies of retraction almost exclusively used PubMed and MEDLINE as their source of data because of the profound implications of maintaining the integrity of the medical and clinical literature.

Table 2. Major aspects of retraction.

Attributes of Retraction	**Findings and References**
Reasons for retraction	Scientific misconduct, irreproducibility, errors (Wager & Williams, 2011); Irreproducibility, unusually high-level of productivity (Budd, et al., 1998; Steen, 2011)
	Misconduct: Identified or presumed; fraud, fabrication, falsification, data plagiarism (Budd, et al., 1998; Neale, et al., 2007; Steen, 2011)
	Errors: Errors in method, data or sample; duplicated publication; text

	plagiarism (Budd, et al., 1998)
Time to retraction (months)	28 months (mean) (Budd, et al., 1998); Fraudulent – 28.41 months (mean), Erroneous – 22.72 months (mean) (Steen, 2011); 28 months (median), Senior researchers implicated – 79 months, junior researcher implicated – 22 months (Trikalinos, et al., 2008); case study (Korpela, 2010)
Post-retraction citations (lag time)	Next calendar year (Pfeifer & Snodgrass, 1990); 1 year after retraction (Budd, et al., 1998); 3 years after (Neale, et al., 2007)
Deliberate or accidental	A higher rate of repeat offenders found in fraudulent papers than erroneous papers (Steen, 2011)
Sources of the literature	PubMed/MEDLINE (Budd, et al., 1998; Neale, et al., 2007; Steen, 2011)

Situation Awareness in a Broader Context

Existing studies of retraction almost exclusively focused on the literature of medicine, where the stake is high in terms of the safety of patients. PubMed and the Web of Science are the major resources used in these studies. Analysts in these studies typically searched for retracted articles and analyzed the content of retraction notices as well as other types of information. Most of these studies appear to rely on labor-intensive procedures with limited or no support of computational and visual analytic tools. Several potentially important questions have not been adequately addressed in part due to such constraints. For example, many eventually retracted articles are highly cited in their fields. It is quite possible that subsequently published articles were unknowingly built on false claims made by a retracted article. In theory, such potentially contaminated articles should be re-examined to ensure that they are valid in light of the retraction. In practice, however, it remains to be a challenging task to assess this type of potential risk timely and systematically. To our knowledge, none of the major scientific content providers such as PubMed, the Web of Science, and Google Scholar readily supports such tasks. Our goal in this article is to provide a better understanding of how tightly a retracted article is interwoven into the scientific literature and what additional actions might be necessary to safeguard the integrity of scientific knowledge.

Tracing the Implicit Impact of Retracted Articles

If an article unknowingly builds on false claims of a retracted article, the new and unsuspicious article may compromise the integrity of the scientific literature. This type of implicit dependency on a retracted article can be highly risky and harmful. Detecting and tracing implicitly infected articles are much harder than identifying officially retracted articles in the literature. Analysts need to assess the potential and actual damage that may be caused by an implicit dependency. Analytic reasoning at this level of granularity is currently beyond the reach of text mining, natural language processing, and science mapping techniques. As the first step towards improving the situation, our goal is to provide visual analytic methods that can assist analysts to identify articles that may implicitly depend on a retracted article.

Existing studies of retraction primarily focused on articles that have been officially retracted, but paid little or no attention to articles that cited a retracted article, or cited an article that cited a retracted article. Over the recent years, tremendous advances have been made in scientometrics (Boyack & Klavans, 2010; Leydesdorff, 2001; Shibata, Kajikawa, & Matsushima, 2007; Upham, Rosenkopf, & Ungar, 2010), science mapping (C. M. Chen, 2006; Cobo, Lopez-Herrera, Herrera-Viedma, & Herrera, 2011; Small, 1999; van Eck & Waltman, 2010), and visual analytics (Pirolli, 2007; Thomas & Cook, 2005). Existing studies of citations to retracted articles have not yet incorporated these relative new and more powerful techniques. Vice versa researchers who have access to the new generation of analytic tools have not applied these tools to the analysis of citation networks involving retracted articles. Our goal is to

demonstrate how visual analytic tools can make such implicit dependencies explicit and easy to detect. In particular, visual analytic tools can be used to reveal the extent to which a single retracted article may be tightly embedded in the literature and visualize the distribution of potentially contaminated articles in a co-citation network.

Suppose an article a_{t_0}, retracted or not, is published at time t_0. A citation path between a_{t_0} and a subsequently published article a_{t_k} can be defined in terms of pairwise citation links $\{a_{t_i}\}$: $a_{t_0} \leftarrow a_{t_1} \leftarrow \cdots \leftarrow a_{t_k}$, where $\leftarrow$ denotes a direct citation, $t_i < t_j$ if $i<j$, and $a_{t_{i+1}}$ has no direct citation link to any of the articles on the path prior to a_{t_i}. The length of a citation path is the number of direct citation links in the path. Existing studies of citations to retracted articles are essentially limited to articles that are one-step away from a retracted article. Longer citation paths originated from a retracted article have not been studied. The retraction of an article is equivalent to the removal of the first article from a potentially still growing path of $a_{t_0} \leftarrow a_{t_1} \leftarrow \cdots \leftarrow a_{t_k}$ because newly published articles may simply follow the article a_{t_k} at the end of the current path without questioning the validity of the potentially risky path as a whole. By *k*-degree post-retraction citation analysis, we introduce a study of such paths formed by *k* pairwise direct citation links as in $a_{t_0} \leftarrow a_{t_1} \leftarrow \cdots \leftarrow a_{t_k}$.

An intuitive way to represent the distribution of articles associated with a retracted article through a citation chain is to visualize a citation or co-citation network and then highlight articles on the citation chain. For example, the retracted article can be depicted in a broader context of how it is co-cited with other articles in the literature and how its interrelationship with the rest of the literature changes over time. More specifically, multiple layers of network visualization can be used to achieve this goal. Each layer of visualization consists of a subset of articles that are *k*-step away from a retracted article. The diffusion process can be shown as the implicit dependency envelop expands across the relevant literature.

Analyzing the Evolution of the Citation Context of Retracted Articles

How is a retracted article cited in subsequently published articles? How did citations differ before and after the retraction? If a new article cited a retracted article, did the authors of the new article know about the retraction? If not, what could be done to improve such situation awareness? If citations to a retracted article are associated with a diverse range of perspectives and opinions, what may be done to assist analysts to analyze and synthesize individual citation instances and form an assessment of the role of the retracted article?

To answer these questions, it is essential to examine the context of a citation to a retracted article and differentiate important arguments made in such citation contexts. A citation context of a cited article is often defined as the sentences that explicitly refer to the article. A more extensive citation context may include more sentences surrounding a citing sentence, for example, *m* sentences before and *n* sentences after, or even the entire paragraph. A retracted article may have been cited by hundreds of subsequently published articles. It would be unrealistic to expect analysts to examine and make sense of a large volume of citation instances without any technical support. An even more challenging task for analysts is to discern emergent patterns from individual citation instances.

Full text articles that cite a retracted article are particularly valuable in developing an understanding of the context of citations to the retracted article. Notable resources of full text articles include PubMed Central (PMC) and arXiv.org. These resources provide a platform for the development of a new generation of visual analytic tools that can analyze and synthesize scientific articles at finer granularity than analytic tools that are limited to the analysis of metadata of scientific articles.

In this article, we demonstrate how citation contexts of retracted articles can provide valuable insights through a topic flow visualization. The citation context of a retracted article evolves over time. The content of its citation context can be characterized in terms of how they change from one year to next. A topic associated with citations to the retracted article in year t_i may evolve in year t_{i+1} in a number of ways.

It may grow stronger, become weaker, or remain the same. Such changes may provide insights into the role of the retracted article in the development of the literature.

In this article, we focus on the issues concerning the context of retracted articles in terms of their interrelationships with other articles in the literature and major themes found in the content of their citation contexts. In particular, our goal is to demonstrate how visual analytic methods and tools can be developed and applied to the study of retracted articles. There are many important issues but, as the first step, we choose to focus on ones that are relatively fundamental.

Method

The focus of our study is on retracted articles that are highly cited in the Web of Science. Retracted articles are potentially harmful to the scientific literature. Highly cited retracted articles could be even more harmful, especially when the bulk of their citations were gathered before it becomes evident to the scientific community that a retraction is necessary. The questions are addressed specifically with these highly cited retracted articles.

Data Collection

We retrieved all the officially retracted articles as follows. In the Web of Science, the title of a retracted article includes a suffix of "*Retracted article*." As of 3/30/2012, there are 1,775 records of retracted articles. Figure 6 depicts the distribution of the 1,775 retracted articles since 1980. The retractions appear to have peaked in 2007 with 254 retracted articles recorded in the Web of Science alone. On the other hand, it might be still too soon to rule out the possibility of more retrospective retractions.

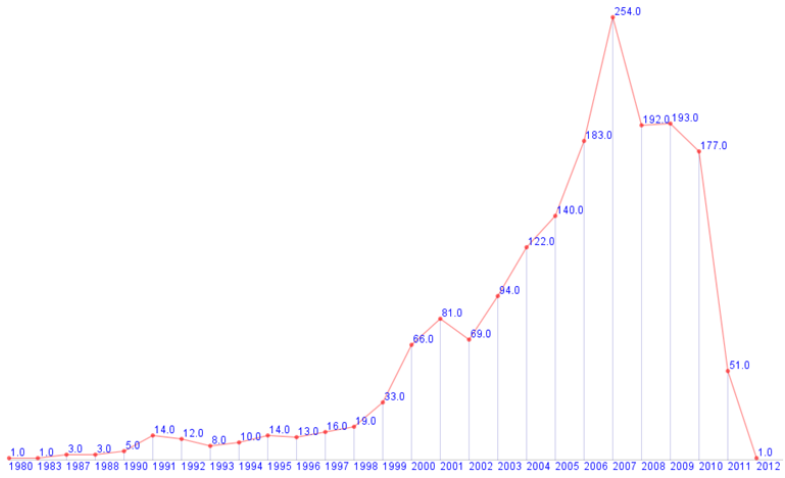


Figure 6. The distribution of 1,775 retracted articles in the Web of Science (as of 3/30/2012). Have we seen the peak yet?

In order to study the scope of the potential contamination, we constructed a larger set of articles that cited the 1,775 retracted articles. Note that citing a retracted article does not necessarily mean that the citing article is contaminated by the flaws of the retracted article. A citing article may well use the retracted article as an example of scientific misconduct. A challenging task for an analyst is to distinguish citations made by authors who may unknowingly build their work on false claims from citations by authors who are fully aware of the problems that led to the retraction of the article they are citing.

Visual Analysis of Retracted Articles and Their Impacts

We constructed the second data set that contains 32,898 articles that cited 1,584 of the 1,775 retracted articles[5], including 22,577 original research articles (68.6%), and 7,179 review articles (21.8%), 1,379 corrections (4.2%), and 1,089 editorials (3.3%). We generated a co-citation network based on 29,756

[5] The fewer number of records is due to the coverage of our institutional subscription of the Web of Science.

original research and review articles between 1998 and 2011 and left out corrections and editorials. Top 50% of the most cited references in each year were chosen to add to the co-citation network with an upper limit of 3,000 references per year.

The topological properties of retracted articles in the visualized network may provide valuable information about the interdependencies between retracted articles and the rest of the literature. For example, if a retracted article is tightly coupled with many references, it tends to appear at the center of a group of cited references. In contrast, if a retracted article is loosely connected to other articles, it may appear in isolation. Generally speaking, it would be much more challenging to eliminate the impact of a well-connected article than an article in isolation. In this type of visualization, the most damaging retracted articles tend to be the ones that are well-connected and highly cited. In other words, one should particularly watch out for those large-sized red dots that are surrounded by many other articles.

Given a particular retracted article, its context in the literature can be defined and represented in several ways. For example, which articles are most often cited together with the eventually retracted article? How are the articles that cited the retracted one distributed in the literature? What have been said about the retracted article when it was cited? The procedure of generating a network visualization regarding a specific retracted article is as follows. Suppose that the retracted article $\mathbf{a_r}$ is cited by $\boldsymbol{M}$ articles in the Web of Science. The $\boldsymbol{M}$ articles also cited a set of many other references – $\boldsymbol{R}$. There are $\boldsymbol{N}$ articles in the Web of Science that cited at least one member of the set $\boldsymbol{R}$. For a chosen retracted article $\mathbf{a_r}$, we retrieved the $\boldsymbol{N}$ articles and generated a co-citation network derived from the $\boldsymbol{N}$ articles. The $\boldsymbol{N}$ articles form a superset of the $\boldsymbol{M}$ articles that cited $\mathbf{a_r}$. Once the co-citation network is visualized, additional layers can be added to the visualization so that articles that cited $\mathbf{a_r}$ directly and indirectly are represented in an extensive citation context.

CiteSpace is a freely available Java application for visualizing structural and temporal patterns in scientific literature (C. Chen, 2006; Chen, Ibekwe-SanJuan, & Hou, 2010). In this study, co-citation networks are generated using CiteSpace. In addition to aggregate citation sentences into clusters at a higher level of abstraction, we further developed a temporal visualization, *topic-flow visualization*, to depict year-by-year flows of topics to assist analytics to discern changes associated with citations to the retracted article. The topic-flow visualization was constructed as follows. First, we group the citation sentences into groups defined by their publication time. Citation sentences made in each year are clustered into topics. Similarities between topics in adjacent years are computed in terms of the overlapping topic terms between them. Topic flows connect topics in adjacent years that meet a user defined similarity threshold.

Results

Time to Retraction

Time to retraction for a retracted article is defined as the length of the duration between the time of its publication and the time of its retraction. The time of publication is routinely available in a bibliographic record of an article. The time of retraction can be found in its bibliographic record in the Web of Science. In the Web of Science, after an article is retracted, its title is updated to indicate the retraction. For example, if the title of a 2010 article contains a phrase "(Retracted article. See vol. 194, pg. 447, 2011)," then we know that the article has been retracted and the year of retraction is 2011. We loaded the retrieved 1,775 records of retracted articles into a built-in relational database of CiteSpace and extracted the year of retraction from the title of each record. The mean of time to retraction is 2.57 years, or 30 months, based on the records of 1,721 retracted articles, excluding 54 records with a missing retraction date. The median of time to retraction is 2 years, i.e. 24 months (see Table 3).

The probability of an eventually retracted article survives retraction at various time points after its publication is analyzed in a survival analysis. Similarity, how likely is a retracted article continuously cited after its retraction is also estimated through a survival analysis. The estimated mean of post-retraction citation is about four years and the median is two years. The estimated mean of citations since

the original publication date is over six years and the median is five years. Based on the median estimates, it tends to take two years to retract an article and another two years to see a significant decrease of citations to the retracted article.

Table 3. Survival analysis of time to retraction, post-retraction citation, and citation since publication.

Survival Event	Mean		Median	
	Estimate	Std. Error	Estimate	Std. Error
Time to retraction	2.578	.006	2.000	.052
Post-retraction citation	4.090	.143	2.000	.146
Citation since publication	6.658	.153	5.000	.187

Figure 7 shows three survival functions over time. The highest solid line is the citation survival function since publication, which shows how likely an eventually retracted article is cited since its publication. The second solid line depicts the post-retraction citation survival function. In other words, it shows how likely a retracted article is cited after retraction. The dashed line is the survival function of retraction, which shows the probability that the article has not been retracted up to that point. According to these survival functions, the majority of retractions took place within the first few years of publication because the dashed line decreases much faster than the other two lines. The probability of surviving retraction more than 4 years is less than 0.2. Post-retraction citations are likely to continue but at a lower and lower rate.

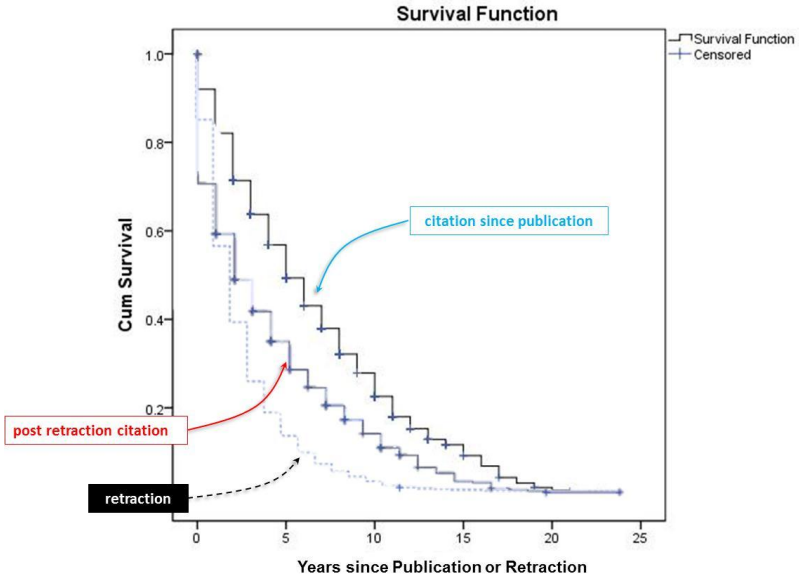


Figure 7. Survival functions of citation since publication (the highest solid line), post-retraction citations (the second solid line), and retraction (the dashed line).

How frequently can a retracted article be cited? Figure 8 plots the average citations received by the retracted articles in the Web of Science. The highest average citation is 113 for retracted articles that were published in 1998. As we will see shortly, this is in part attributed to a 1998 article, which is the most highly cited retracted article ever. A total of 36,218 articles in the Web of Science cited members of the set of 1,775 retracted articles 39,557 times, excluding self-citations. On average, each retracted article is cited 22.29 times. The h-index of this set of retracted articles is 88, which means that 88 of the retracted articles have been cited 88 times or more. These citation statistics indicate that retracted articles may have a considerable degree of impact on the scientific literature. Retracting directly involved articles may not be effective enough to stop a continuous spread of a potentially harmful impact.

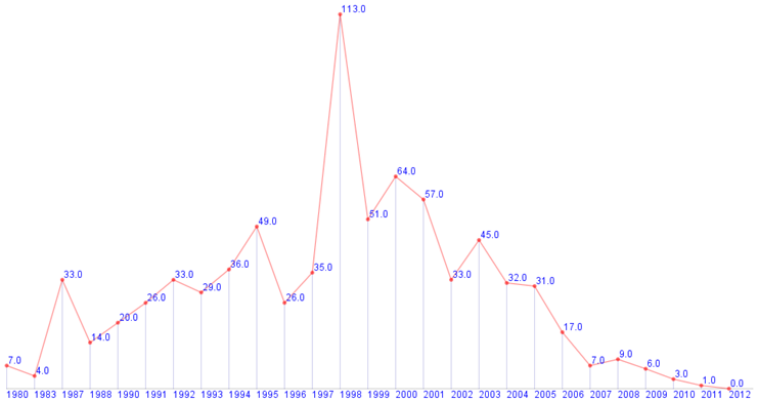


Figure 8. The distribution of the average citation counts of 1,775 retracted articles.

Highly Cited Retracted Articles

Table 4 lists the citation counts of the 10 most highly cited retracted articles in the Web of Science. Each of the ten articles has been cited hundreds of times. 740 articles cited the 1998 *Lancet* paper by Wakefield et al., the first article on the list, whereas 366 articles cited the 10th article on the list. Three papers on the list were published in *Science* and two in *Lancet*. In the rest of the article, we will primarily focus on these high-profile retracted articles in terms of their citation contexts at both macroscopic and microscopic levels.

Table 4. The 10 most highly cited retracted articles.

Citations	Lead Author	Publication —Retraction	Title (Retraction Notice)	Journal
740	Wakefield, AJ	1998—2010	Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children (See vol 375, pg 445, 2010)	LANCET
727	Reyes, M	2001—2009	Purification and ex vivo expansion of postnatal human marrow mesodermal progenitor cells (See vol. 113, pg. 2370, 2009)	BLOOD
659	Fukuhara, A	2005—2007	Visfatin: A protein secreted by visceral fat that mimics the effects of insulin (See vol 318, pg 565, 2007)	SCIENCE
618	Nakao, N	2003—2009	Combination treatment of angiotensin-II receptor blocker and angiotensin-converting-enzyme inhibitor in non-diabetic renal disease (COOPERATE): a randomised controlled trial (See vol. 374, pg. 1226, 2009)	LANCET
512	Chang, G	2001—2006	Structure of MsbA from E-coli: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters (See vol 314, pg 1875, 2006)	SCIENCE
492	Kugler, A	2000—2003	Regression of human metastatic renal cell carcinoma after vaccination with tumor cell-dendritic cell hybrids (See vol. 9, p. 1221, 2003)	NATURE MEDICINE
433	Rubio, D	2005—2010	Spontaneous human adult stem cell transformation (See vol. 70, pg. 6682, 2010)	CANCER RESEARCH
391	Gowen, LC	1998—2003	BRCA1 required for transcription-coupled repair of oxidative DNA damage (See vol 300, pg 1657, June 13 2003)	SCIENCE
375	Hwang, WS	2004—2006	Evidence of a pluripotent human embryonic stem cell line derived from a cloned blastocyst (See vol 311, pg 335, 2006)	SCIENCE
366	Makarova, TL	2001—2006	Magnetic carbon (See vol 440, pg 707, 2006)	NATURE

In order to identify a meaningful context of retracted articles, we constructed a network of co-cited scientific publications that is broad enough to represent the underlying knowledge structure. A co-citation network of scientific publications consists of scientific publications, or cited references, as nodes. A co-

citation link between two nodes represents how often the two nodes are cited together in subsequent publications. The relevant literature on using co-citation networks to represent the intellectual structure of an underlying subject domain or a discipline.

Retracted Articles in a Co-Citation Network

The overall co-citation network, containing 7,217 cited references and 155,391 co-citation links, was visualized as a base map of the context of the 1,584 retracted articles. Next, the 1,584 retracted articles were projected on top of the base map (See Figure 9).

Each dot in the visualization represents a reference cited by the 29,756 articles. The dots in red indicate articles that were retracted eventually. Lines between dots are co-citation links. The color of a co-citation link is the earliest time a co-citation between two articles was made. The earliest links are colored in blue; more recent links in yellow and orange. The size of a dot, or a disc, is proportional to the citation counts of the corresponding cited article. The top 10 most cited retracted articles are labeled in the visualization.

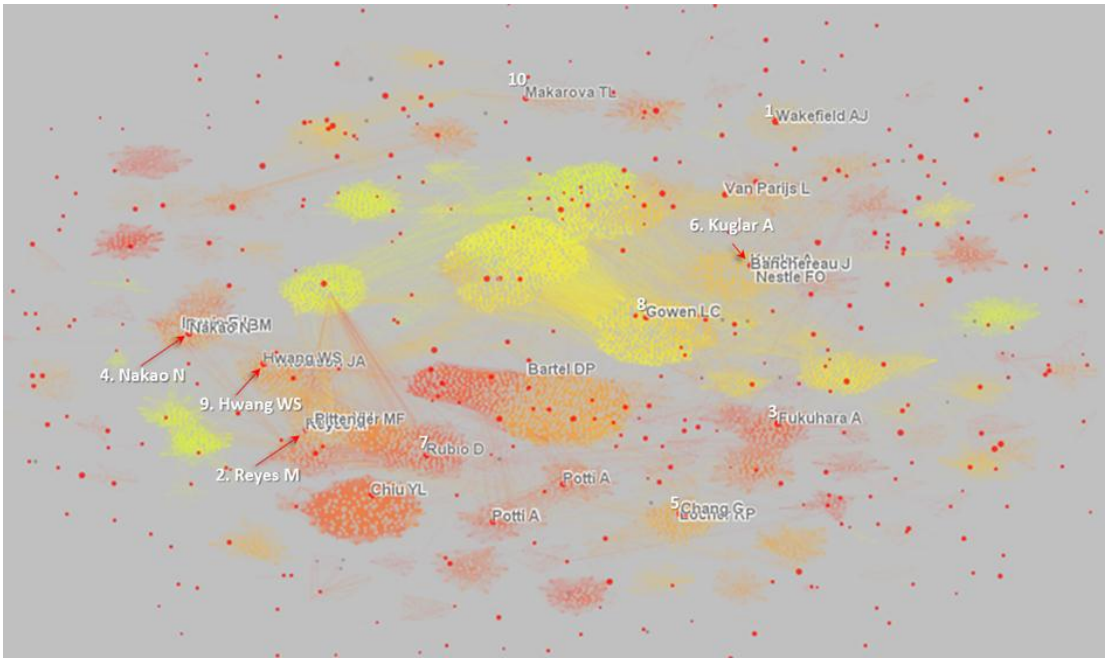


Figure 9. An overview of co-citation contexts of retracted articles. Each dot is a reference of an article. Red dots indicate retracted articles. The numbers in front of labels indicate their citation ranking. Potentially damaging retracted articles are in the middle of an area that otherwise free from red dots.

Figure 10 shows a more detailed local view of the contextual overview, highlighting several articles associated with some of the most high-profile retraction cases in the recent history of science. The article by Nakao N et al. on the left, for example, was originally published in the *Lancet* in 2003. It reported that combination therapy of an agiotensin-II receptor blocker and angiotensin-converting-enzyme (ACE) inhibitor in non-diabetic renal disease was superior to an ACE inhibitor alone. The article was retracted in 2009 after the lead author was engaged in serious scientific misconduct. Many patients were adversely affected by the publication. Its position on a densly connected island of other articles indicates its relevant to a significant topic. The visualization also shows the positions of retracted articles by Hwang WS, slightly to the right, and Potti A, at the lower right corner. They have similar citation context profiles. Interconnected citation contexts of multiple retracted articles are also areas where analysts should pay attention.

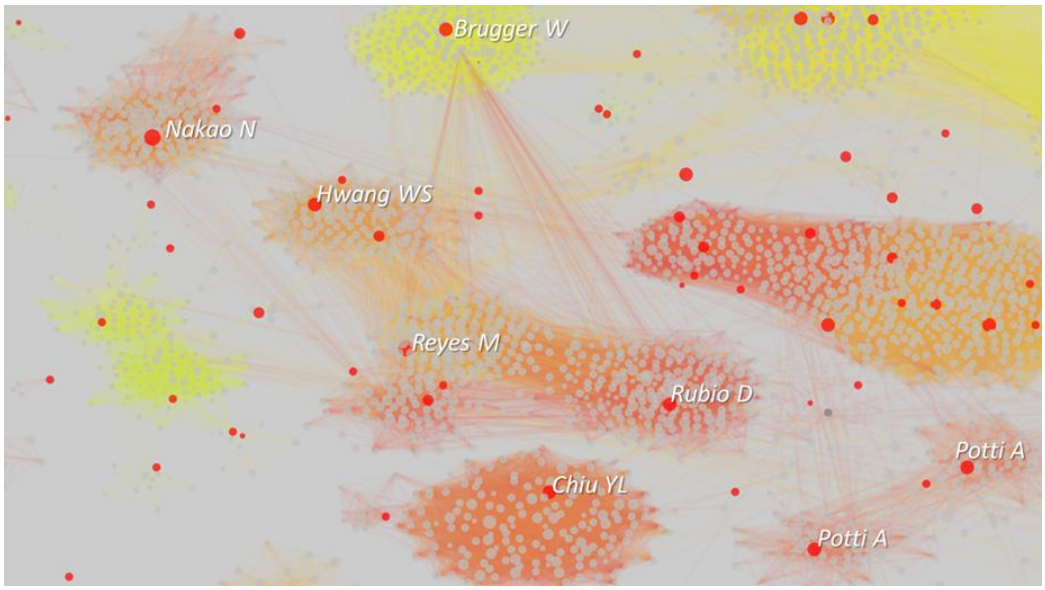


Figure 10. Red dots are retracted articles. Labeled ones are highly cited. Clusters are formed by co-citation strengths.

Figure 11 shows an extensive representation of the citation context of the retracted 2003 article by Nakao et al. First, 609 articles that cited the Nakao paper were identified in the Web of Science. Next, 9,656 articles were retrieved because they have at least one common references with the 609 direct citing articles. Top 6,000 most cited references per year between 2003 and 2011 were chosen to form a co-citation network of 27,905 references and 2,162,018 co-citation links. The retracted Nakao paper is shown as the black dot in the middle of the map. The red dots are 340 direct citers of the total of 609 available in the Web of Science. The cyan dots share common references with the direct citers, not necessarily the retracted article. The labels are the most cited articles in this topic area, which are not retracted articles themselves.

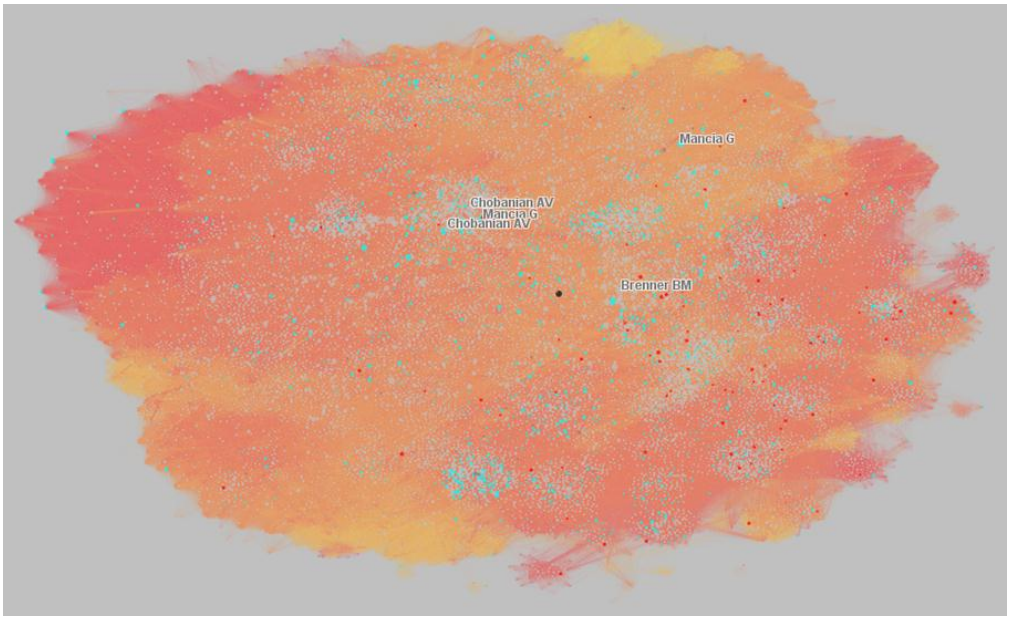


Figure 11. An extensive citation context of a retracted 2003 article by Nakao et al. The co-citation network contains 27,905 cited articles between 2003 and 2011. The black dot in the middle of the dense network represents the Nakao paper. Red dots represent 340 articles that directly cited the Nakao paper (there are 609 such articles in the Web of Science). Cyan dots represent 2,130 of the 9,656 articles that bibliographically coupled with the direct citers.

Visualizing the extensive citation context of a retracted article is potentially valuable for analysts and scientists to estimate more accurately the scope of a retracted article's influence on the scientific literature. This type of visual analytic tools can be used to facilitate otherwise complex and intangible situation awareness tasks involving the retraction of a single article.

Broader Context of High-Profile Retracted Articles

The potential damage of a retracted article depends on how tightly it is interwoven with the rest of the literature. If the retracted article is part of a fast-growing area of research, it would be more damaging than a retracted article from a slow-moving area. In the following example, we analyze the temporal properties associated with the top 10 retracted articles in order to identify the growth of corresponding research areas of these retracted articles.

We retrieved 29,756 bibliographic records of the types of Article and Review only from the Web of Science. These articles shared at least one reference in common with the top 10 retracted articles. Citations made by the 29,756 articles therefore provide an adequate representation of the research areas where these retracted articles belong to. CiteSpace was used to generate a synthesized co-citation network based on individual co-citation networks formed from 1990 to 2011 using the top 30 most cited references each year. Although more references can be sampled per year for the analysis, our focus is on the temporal patterns that might be related to the major research areas where retracted articles are found.

Figure 12 shows a timeline visualization of the 37 clusters of co-cited references. Each cluster represents a research specialty. Circles with blue labels are 5 retracted articles from the top 10 list. Citation bursts are abrupt increases of citations. A citation burst indicates a possible hot research area. Citation bursts are shown as red rings in the timeline visualization. For example, Cluster #1 visfatin has a few big circles with red rings of citation bursts. In particular, it contains a retracted 2005 article by Fukuhara et al. Similarly, Cluster #12 contains a series of articles with citation bursts and a retracted article by Nakao et al. Cluster #18, labeled as mesenchymal stem cell, has a string of articles with citation bursts, including a retracted 2001 article by Reyes et al. The considerable amount of clusters with significant citation bursts suggest that many of these retracted articles are from active and significant research areas. Therefore, purging the negative impact of these high-profile retracted articles is far more challenging than retracting these articles alone because it may become necessary to re-examine the entire research area and re-establish the credibility of research involving many other innocent researchers in the worst-case scenario.

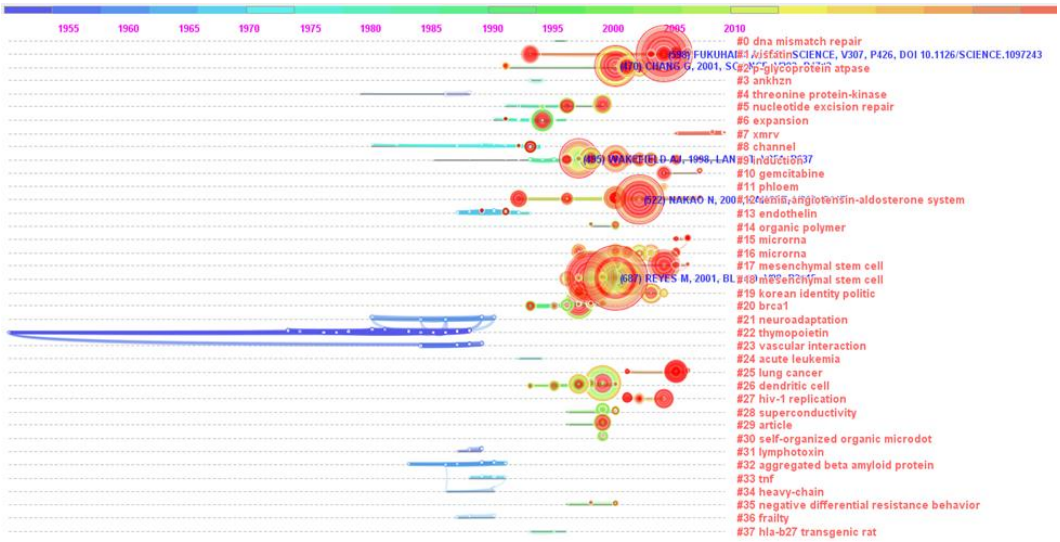


Figure 12. Research areas where the top 10 retracted areas belong to. Red rings indicate citation bursts, indicating vibrant research areas. Blue labels indicate retracted articles.

Figure 13 shows clusters with 5 or more co-cited references along with three sets of labels chosen from titles of citing articles that formed these clusters. The higher the silhouette value a cluster has, the more homogenous the cluster is. The mean year of publication is the average year in which member articles were published.

Cluster...	Size	Silhouette	mean...	Top Terms (tf*idf weighting)	Top Terms (log-likelihood ratio, p-level)	Terms (mutual information)
9	43	0.186	1998	(14.91) antibody; (14.91) vaccine; (14.16) c...	induction (49.28, 1.0E-4); evolution (47.87, ...	adult blood vessel formation
22	31	0.912	1982	(21.13) acetylcholine-receptor; (20.84) thy...	thymopoietin (165.13, 1.0E-4); thymopentin ...	synthesis
20	29	1	1997	(22.46) dna damage response; (21) mam...	brca1 (230.81, 1.0E-4); dna damage respo...	accumulation
8	25	1	1992	(18.42) asn; (18.19) potassium channel; (1...	channel (181.16, 1.0E-4); asn (164.47, 1.0E...	membrane protein
13	22	1	1991	(20.14) endothelin receptor; (19.48) endoth...	endothelin (106.69, 1.0E-4); endothelin rec...	salt-treated rat
6	17	1	1994	(18.14) peripheral blood; (17.4) peripheral ...	expansion (142.97, 1.0E-4); peripheral bloo...	transduction
32	15	0.989	1990	(11) kinin-forming cascade; (11) aggregate...	aggregated beta amyloid protein (41.42, 1.0...	epilepsy
16	12	0.957	2002	(12.03) dna methyltransferase; (11.77) mic...	microrna (36.27, 1.0E-4); small non-coding ...	paradox
7	9	1	2009	(18.86) contamination; (18.04) xmrv; (15.67...	xmrv (179.92, 1.0E-4); chronic fatigue syndr...	xpr1 receptor
18	9	0.872	2000	(14.44) stromal cell; (13.56) mesenchymal ...	mesenchymal stem cell (78.01, 1.0E-4); bo...	chondrogenic effect
1	8	1	2004	(28.15) visfatin; (18.86) adipokine; (18.65) i...	visfatin (521.12, 1.0E-4); insulin resistance ...	cause
12	8	1	2001	(18.52) angiotensin; (17.25) renin-angioten...	renin-angiotensin-aldosterone system (99....	oxidative stres
23	8	0.985	1987	(9.68) vascular interaction; (9.68) mammali...	vascular interaction (34.49, 1.0E-4); mamm...	human bone-marrow-derived mesenchyma...
21	7	0.957	1987	(10.15) neuroadaptation; (10.15) desensiti...	neuroadaptation (35.14, 1.0E-4); desensitiz...	neuroadaptation
26	7	1	1998	(18.24) dendritic cell; (14.43) vaccination; (...	dendritic cell (98.84, 1.0E-4); cancer (59.68,...	autologous t lymphocyte
33	7	0.793	1990	(12.58) tnf; (10.86) tumor-necrosis-factor; (...	tnf (50.52, 1.0E-4); tnf-alpha (20.12, 1.0E-4);...	tumor necrosis factor-alpha
5	6	1	1995	(14.25) repair; (13.81) nucleotide excision r...	nucleotide excision repair (72.96, 1.0E-4); c...	oxidative stres
2	5	1	2000	(13.58) p-glycoprotein; (11.46) glycoprotein...	p-glycoprotein atpase (19.96, 1.0E-4); struct...	relevance
25	5	1	2005	(16.07) classification; (16.06) lung; (15.47) ...	lung cancer (110.57, 1.0E-4); non-small cell...	adenocarcinoma
27	5	1	2003	(16.44) restriction factor; (15.54) restriction;...	hiv-1 replication (60.38, 1.0E-4); apobec3 cy...	hiv-1 persistence
28	5	0.894	1999	(15.75) superconductivity; (15.67) mgb; (14...	superconductivity (155.64, 1.0E-4); superco...	molecular design
31	5	1	1989	(7.81) hematopoietic cytokine; (7.81) lymph...	lymphotoxin (23.05, 1.0E-4); human fibrobla...	tumor-necrosis-factor-alpha

Figure 13. Clusters with 5 or more co-cited references.

The Wakefield Article

We inspected the citations to the Wakefield article in order to have a better understanding of how a retracted article may affect the scientific literature. The *Lancet* partially retracted the 1998 article in 2004 and retracted fully in 2010. The Lancet's retraction notice in February 2010 noted that several elements of the 1998 paper are incorrect, contrary to the findings of an earlier investigation, and that the paper made false claims of an "approval" of the local ethics committee.

The Wakefield article was cited by 740 publications at the time of writing, including some high-impact citers. The two of the most prominent citers have 384 and 360 citations, respectively. The third high-impact citer is a 1999 article (Taylor et al., 1999), which has 296 citations. The 740 direct citers were cited by 6,600 articles in the Web of Science, which in turn cited 12,612 references. A citation burst of 0.05 was detected for the Wakefield article, indicating that the article had drawn a considerable amount of attention during a short period of time. Its citation counts peaked in 2002.

Figure 14. The citation history of the Wakefield paper. Its citations peaked in 2002. It was partially retracted in 2004 and fully retracted in 2010.

The citation history of the Wakefield paper is shown in Figure 14. Its citations peaked at 53 citations in 2002 and dropped to 35 a year after its 2004 partial retraction. Then it increased to 42 before dropped down to 31 in 2008. Interestingly, the paper was cited 47 times in 2011 and it was its highest annual citation count since 2003.

In order to identify the nature of the majority of its citations, we further studied the sentences that contain the Wakefield paper as a reference. We refer to these sentences as citing sentences. Citing sentences were extracted from full text articles retrieved from Elsevier's Content Syndication (ConSyn). ConSyn contains 3,359 scholarly journals and 6,643 non-serial titles. Since the Wakefield paper is concerned with a claimed causal relation between a combined MMR vaccine and autism, we searched for full text journal articles on autism and vaccine in ConSyn and found 1,250 relevant full text articles. Among the 1,250 full text articles, 156 explicitly cited the Wakefield paper in 706 distinct sentences. These sentences were grouped into 69 clusters by the Lingo clustering method in Carrot2, an open-source framework for building search clustering engines[6].

Figure 15 shows a FoamTree visualization of the 69 clusters of the 706 sentences that cited the Wakefield paper. Clusters with the largest areas represent the most prominent clusters of phrases used when researchers cited the Wakefield paper. Notably, *inflammatory bowel disease*, *mumps and rubella*, and *association between MMR vaccine and autism* are the central topics of the citations. These topics characterize the key issues surrounding the retracted *Lancet* paper.

We did not differentiate positive and negative citations in the present study. Identifying the orientation of an instance of citation from a citation context, for example, the citing sentence and its surrounding sentences, is a very challenging task even for an intelligent reader because the position of the argument becomes clear only when a broader context is taken into account, for example, after reading the entire paragraph in many cases.

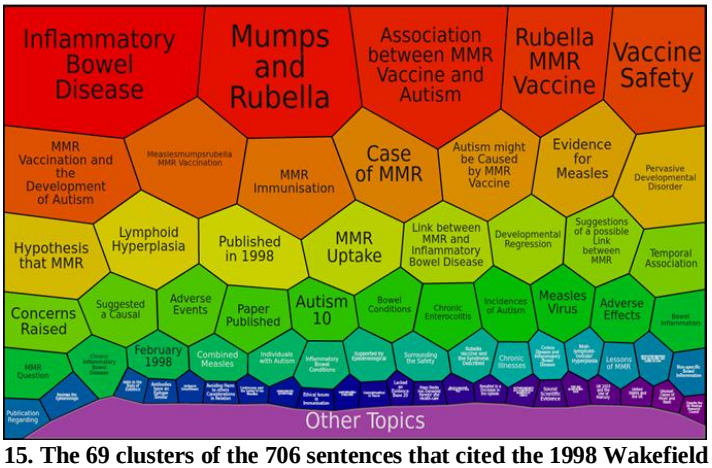


Figure 15. The 69 clusters of the 706 sentences that cited the 1998 Wakefield paper.

Topic-flow visualization characterizes each topic as convergent and divergent as well as steady topics. A convergent topic in a particular year is defined in terms of the number of related topics in the previous year. The convergent topic sums up elements from multiple previously separated topics. Topics labeled in Figure 16 are examples of convergent topics. In 1999, the topic of Rubella MMR Vaccination is highlighted by an explicit label because it is associated with several distinct topics in 1998. In 2004, the year Lancet partially retracted the Wakefield paper, the prominent convergent topic was Developmental Disorders. The visualization shows that numerous distinct topics in 2003 were converged into the

[6] http://project.carrot2.org/

convergent topic in 2004. We expect that this type of topic-flow visualizations can enable new ways of analyzing and studying the dynamics of topic transitions in specific citations to a particular article.

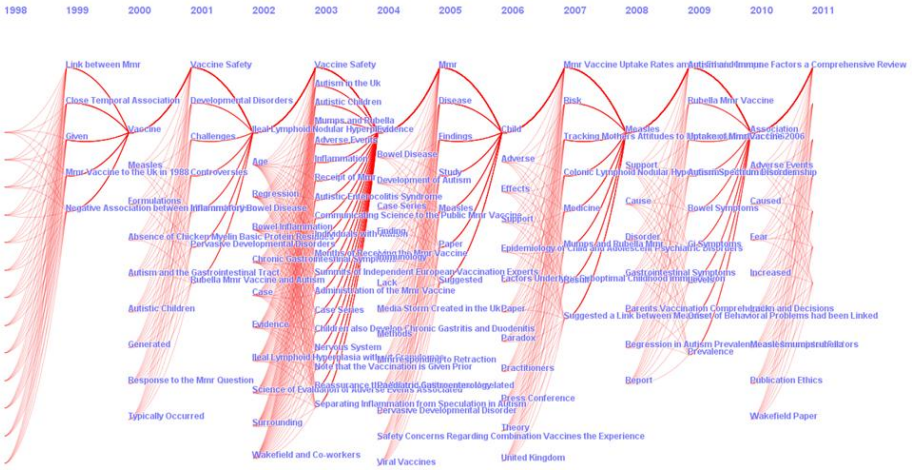


Figure 16. A topic-flow visualization of specific citations to the Wakefield paper. Convergent topics are identified in terms of a relatively high in-degree of flows from previous years, for example, the topic of Vaccine in 2000 and the topic of Evidence in 2004. Topics in each year are vertically aligned by the size of topic.

In contrast to the notion of convergent topics in topic-flow visualization, divergent topics in a given year are characterized in terms of how many related topics in the subsequent year. As shown in Figure 17, divergent topics in 2000 include Measles and Autistic Children. The divergent topic found in 2002 is Ileal-lymphoid-nodular hyperplasia, which is part of the title of the Wakefield paper. Convergent and divergent topics together provide a rich set of information about the dynamics of topic transitions over time, which in turn provides a new layer of information that is more specific than the macroscopic patterns and easier to discern than examining individual sentences alone.

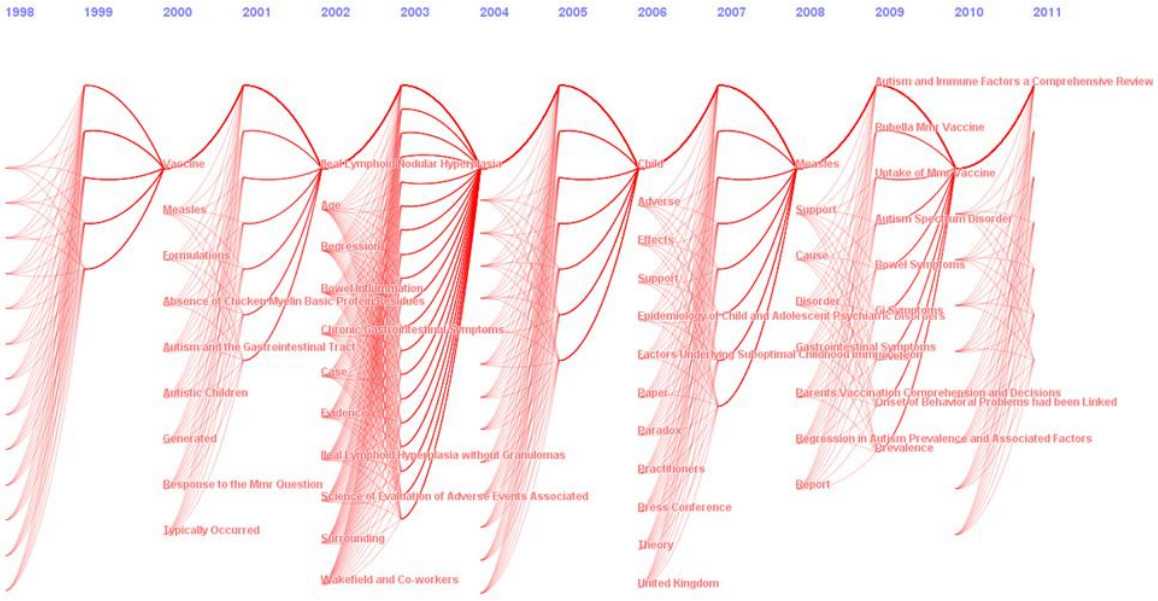


Figure 17. Divergent topics in a topic-transition visualization of the Wakefield paper.

Topic-flow visualizations provide a novel interface for analysts to explore specific sentences that cite a particular article. In this article, the focus is on retracted articles. In general, this approach is applicable to

a broad range of articles, for example, for studying citation patterns before and after a citation burst. The visual interface can also facilitate a drill-down analysis of individual sentences or other types of evidence in full text. Convergent and divergent topics are valuable for the development of a general framework for detailed studies of the history of a retracted article. Table 5 lists examples of sentences that cited the 1998 Lancet paper by Wakefield et al. For example, as early as 1998, researchers were concerned about the lack of sound scientific evidence to support the claimed association between MMR vaccine and inflammatory bowel disease. The adverse impact on MMR uptake is also evident in these citation sentences. Many more analytic tasks may become feasible with this type of text and pattern-driven analyses at multiple levels of granularity.

Table 5. Specific sentences that cite the eventually retracted 1998 Lancet paper by Wakefield et al.

Year of Citation	Reference	Sentence
1998	1	The report by Andrew Wakefield and colleagues confirms the clinical observations of several paediatricians, including myself, who have noted an association between the onset of the autistic spectrum and the development of disturbed bowel habit.
1998	1	Looking at the ages of the children in Wakefield's study, it seems that most of them would have been at an age when they could well have been vaccinated with the vaccine that has since been withdrawn.
1998	1	We are concerned about the potential loss of confidence in the mumps, measles, and rubella (MMR) vaccine after publication of Andrew Wakefield and colleagues' report (Feb 28, p 637), in which these workers postulate adverse effects of measles-containing vaccines.
1998	1	We were surprised and concerned that the Lancet published the paper by Andrew Wakefield and colleagues in which they alluded to an association between MMR vaccine and a nonspecific syndrome, yet provided no sound scientific evidence.
2001	34	In 1998, Wakefield et al.[34] have published a second paper including two ideas: that autism may be linked to a form of inflammatory bowel disease and that this new syndrome is associated with measles–mumps–rubella (MMR) immunization.
2007	5	Vaccine scares in recent years have linked MMR vaccination with autism and a variety of bowel conditions, and this has had an adverse impact on MMR uptake [5].
2007	5	When comparing MMR uptake rates before (1994-1997) and after (1999-2000) the 1998 Wakefield et al. article [5] it is seen that prior to 1998 Asian children had the highest uptake.
2010	2	This addresses a concern raised by a now-retracted article by Wakefield et al. and adds to the body of evidence that has failed to show a relationship between measles vaccination and autism (1,2).

In order to demonstrate the generic value of the topic-transition visualization in the study of citation context, we include a topic-transition diagram associated with citations to another high-profile retracted article, the fraudulent 2004 Science paper on cloning stem cells by Hwang et al. The 2004 paper was retracted in 2006[7] (Kennedy, 2006). This case is analyzed in several studies (Kakuk, 2009). According to the editor of Science (Kennedy, 2006), Seoul National University's investigation committee concluded that the authors of the Hwang paper (and another paper that appeared in 2005 Science) engaged in research misconduct and that the papers contain fabricated data. The results reported in these papers are deemed to be invalid. We were able to retrieve 1,068 full text articles in Elsevier's ConSyn by formulating a query with title words of the Hwang paper, i.e., pluripotent human embryonic stem cell cloned blastocyst. 72 of the full text articles cited Hwang's 2004 Science paper. Figure 18 shows a topic-transition diagram of topics from citation sentences to the eventually retracted paper. Both convergent and

[7] http://en.wikipedia.org/wiki/Hwang_Woo-Suk

divergent topics are labeled in this view. The paper was retracted in 2006. The convergent topic of Nuclear-Transfer is prominent in the diagram. Table 6 includes examples of citations to the Hwang paper.

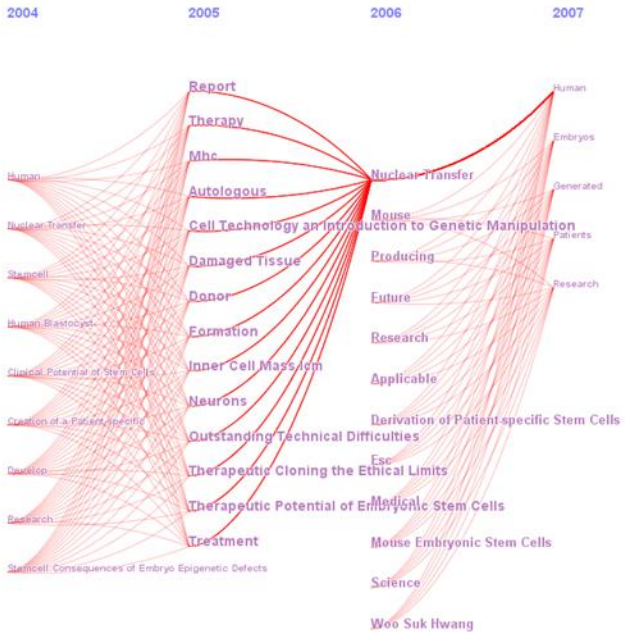


Figure 18. Convergent and divergent topics in specific citation sentences to the Hwang paper. Convergent topics are the ones to which many previous topics converged. Divergent topics are the ones that branch out to many new topics.

Table 6. Citation sentences to the 2004 Hwang paper.

Year	Reference	Title of Citing Article	Sentence
2004	17	Stem-cell consequences of embryo epigenetic defects	This recent approach, by Woo Suk Hwang and colleagues, produced a single cell-line from 242 oocytes recovered from 16 donor women.
2006	3	Plagiarism Digging to the root of the problem	There is the infamous case in which Science had to retract two articles by Woo Suk Hwang et al. [2, 3] because of falsification of data.
2007	4	Acquiring human embryos for stem-cell research	Early reports of success in human therapeutic cloning [4, 5], were retracted as false, but the scientific potential, for instance to create new tissues or organs for disabled patients, remains.

Discussions

Our study aims to raise the awareness of the increasing prevalence of retractions in scientific literature and the complexity and challenges associated with minimizing direct and indirect damages caused by retracted articles. For instance, we have shown that the rate of retraction is increasing based on the numbers of retraction notices and the total number of publications found in PubMed. We found that it takes about 2 years on average to retract an article and another two years to see a substantial decrease of citations to the retracted article.

We have shown that many retracted articles have been cited hundreds of times. Visualizations of co-citation networks demonstrate that many retracted articles are deeply interwoven with the rest of literature. Retracting these articles alone is less likely to purge the literature adequately because the current practice does not systematically re-examine articles that cited these retracted articles and new articles may unknowingly cite a chain of such articles.

More importantly, verifying the validity of articles on citation chains becomes increasingly challenging as new publications are added to the literature and their validity may be taken for granted because they are not directly involved in any retractions. New mechanisms for checking plagiarism, duplication, and indirect citations to retracted articles in new manuscripts should be considered as an integral part of a manuscript management workflow. A survival analysis in this article has shown that an article may remain in the literature for as long as ten years or more before its ultimate retraction. Our study may raise the awareness of the potential risk of articles associated with retracted articles through high-order, indirect chains of citations.

We have demonstrated with visualization and science mapping techniques that many retracted articles are highly cited as part of vibrant lines of research. In other words, these retracted articles are potentially more dangerous than retracted articles in less active areas of research, especially when no effective tools are readily available to track down closely related articles. We recommend that the study of scientific literature should be done routinely such that retracted articles and closely related articles can be identified timely.

We have demonstrated how a visual analytics approach can be used to facilitate the study of the role played by retracted articles. For instance, topic-flow visualizations derived from citation sentences can bridge the cognitive and conceptual gap between macroscopic patterns and microscopic individual instances. The topic flow of citation sentences can be used to identify convergent and divergent topics, which will enable analysts to discern the dynamics of topic transitions associated with the role played by a retracted article.

In comparison with the existing studies of retracted articles, our study has made the following contributions:

1. We have shown that the rate of retraction is increasing based on the numbers of retraction notices and the total number of publications found in PubMed.
2. We have shown that many retracted articles are highly cited with hundreds of citations and they are often part of active areas of research. These findings underline the urgency of identifying the extent they pose a threat to the credibility of the literature.
3. Retracting these articles alone is unlikely to eliminate the risk completely from the scientific literature because new publications may still unknowingly extend a citation trial originated from a retracted article.
4. More importantly, verifying the validity of such citation trails is likely to become increasingly challenging as more publications become attached to the trails and researchers may take their validity for granted. New visual analytic tools provide a useful support.

At a disciplinary level, we expect that our study can draw attention of both science mapping researchers and a broad range of users who may concern with the integrity of scientific knowledge such that a new generation of visual analytics tools can take retracted articles into account in addition to the traditional focus on the literature of science. Analysts, policy makers, and regulatory authorities are among the potential users to benefit from a broader view of the scientific literature and an increased awareness of the challenges and ways to maintain and safeguard the integrity of science. We expect that our study may stimulate a new generation of analytical features of science mapping and visual analytics systems.

Conclusions

The contributions of our study include 1) demonstrating the extent to which retracted articles could impact on subsequent research; specifically, retracted articles are not isolated in co-citation networks and sentence-level contexts where eventually retracted articles were cited so as one can better estimate the negative impact; 2) demonstrating how these techniques taken together offer a visual analytic approach to gain insights into the interrelationship between retracted articles and the rest of the scientific literature. These contributions are significant with practical implications.

The perceived risk introduced by retracted articles alone is the tip of an iceberg. Many retracted articles are highly cited as part of a fast-moving field of research. It is essential to raise the awareness that much of the potential damages introduced by a retracted article are hidden and lasting well beyond the retraction. The original attention drawn to a retracted article may be lost after generations of subsequent citations.

New tools and services are needed to enable researchers and analysts better deal with the increasing prevalence of retractions and safeguard the integrity of scientific literature. In particular, tools are needed to verify the status of a citation genealogy to ensure that the current status of the origin of the genealogy is clearly understood. Such tools should become part of the workflow of journal editors and publishers as well as individual scientists.

From a visual analytic point of view, it is essential to bring in more techniques and tools that can support analytic and sense making tasks from the dynamic and unstructured information and allow analysts and researchers to move back and forth freely across multiple levels of analytic and decision making tasks. The ability of trailblazing evidence and arguments through an evolving space of knowledge is a critical step for the creation of scientific knowledge and maintaining a trustworthy documentation of the collective intelligence.

Acknowledgement

The work would not be possible without the citation records of scientific publications from Thomson Reuter's Web of Science and full text articles from Elsevier's Content Syndication (ConSyn). CC and TS would like to acknowledge the support of a sponsored research project by Pfizer. ZH is a visiting doctoral student at the College of Information Science and Technology, Drexel University. His visit is sponsored by the China Scholarship Council.

References

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389-2404.

Budd, J. M., Sievert, M., & Schultz, T. R. (1998). Phenomena of retraction: Reasons for retraction and citations to the publications. *JAMA, 280*, 296-297.

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

Cobo, M. J., Lopez-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science Mapping Software Tools: Review, Analysis, and Cooperative Study Among Tools. [Review]. *Journal of the American Society for Information Science and Technology, 62*(7), 1382-1402.

Kakuk, P. (2009). The legacy of the Hwang case: research misconduct in biosciences. *Sci Eng Ethics, 15*, 545-562.

Kennedy, D. (2006). Editorial Retraction. *Science, 311*(5759), 335.

Korpela, K. M. (2010). How long does it take for scientific literature to purge itself of fraudulent material? The Breuning case revisited. *Curr Med Res Opin, 26*, 843-847.

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

Naik, G. (2011, August 10, 2011). Mistakes in scientific studies surge. *The Wall Street Journal* Retrieved 3/16/2012, 2012, from http://online.wsj.com/article/SB10001424052702303627104576411850666582080.html

Neale, A. V., Northrup, J., Dailey, R., Marks, E., & Abrams, J. (2007). Correction and use of biomedical literature affected by scientific misconduct. *Sci Eng Ethics, 13*, 5-24.

Pfeifer, M. P., & Snodgrass, G. L. (1990). The continued use of retracted, invalid scientific literature. *Journal of the American Medical Association, 263*, 1420-1423.

Pirolli, P. (2007). *Information Foraging Theory: Adaptive Interaction with Information*. Oxford, England: Oxford University Press.

Service, R. F. (2002). Bell Labs fires star physicist found guilty of forging data. *Science, 298*, 30-31.

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.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799-813.

Sox, H. C., & Rennle, D. (2006). Research misconduct, retraction, and cleansing the medical literature: lessons from the Poehlman case. *Annals of Internal Medicine, 144*, 609-613.

Steen, R. G. (2011). Retractions in the scientific literature: Do authors deliberately commit research fraud? *J Med Ethics, 37*, 113-117.

Taylor, B., Miller, E., Farrington, C. P., Petropoulos, M. C., Favot-Mayaud, I., Li, J., et al. (1999). Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association. [Article]. *Lancet, 353*(9169), 2026-2029.

Thomas, J., & Cook, K. (2005). *Illuminating the Path, the Research and Development Agenda for Visual Analytics*: IEEE CS Press.

Trikalinos, N. A., Evangelou, E., & Ioannidis, J. P. A. (2008). Falsified papers in high-impact journals were slow to retract and indistinguishable from nonfraudulent papers. *Journal of Clinical Epidemiology, 61*, 464-470.

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

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. [Article]. *Scientometrics, 84*(2), 523-538.

Wager, E., & Williams, P. (2011). Why and how do journals retract articles? An analysis of Medline retractions 1988-2008. *J Med Ethics, 37*, 567-570.

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., et al. (1998). Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children (Retracted article. See vol 375, pg 445, 2010). *The Lancet, 351*(9103), 637-641.