Identifying Transformative Scientific Research

Yi-hung Huang[1,2], Chun-Nan Hsu[2,3], and Kristina Lerman[3]
[1]Department of Computer Science, National Taiwan University, Taipei 106, Taiwan
[2]Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
[3]USC Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292, USA
Email:lerman@isi.edu

***Abstract*—Transformative research refers to research that shifts or disrupts established scientific paradigms. Notable examples include the discovery of high-temperature superconductivity that disrupted the theory established 30 years ago. Identifying potential transformative research early and accurately is important for funding agencies to maximize the impact of their investments. It also helps scientists identify and focus their attention on promising emerging works. This paper presents a data-driven approach where citation patterns of scientific papers are analyzed to quantify how much a potential challenger idea shifts an established paradigm. The key idea is that transformative research creates an observable disruption in the structure of "information cascades," chains of references that can be traced back to the papers establishing some scientific paradigm. Such a disruption is visible soon after the challenger's introduction. We define a *disruption score* to quantify the disruption and develop an algorithm to compute it from a large citation network. Experimental results show that our approach can successfully identify transformative scientific papers that disrupt established paradigms in Physics and Computer Science, regardless of whether the challenger paradigm is an instant hit or a classic whose contribution is formally recognized with a Nobel Prize decades later.**

I. Introduction

Transformative research refers to research driven by ideas that lead to emerging concepts, approaches, and/or new subfields of research that shifts or disrupts an established scientific paradigm [1]. Thomas Kuhn's influential book titled *The Structure of Scientific Revolutions* [2] describes the progress of science as non-linear, propelled by "paradigm shifts" in which scientists' world-views, or paradigms, are altered dramatically by a new discovery, theory, or methodology. Ability to systematically identify transformative research has numerous benefits, including helping funding agencies establish funding priorities, allowing individual scientists to better keep up with important new research, and translating transformative research faster into practice.

However, identifying transformative ideas is not easy. The process by which such ideas are recognized and accepted by the scientific community is affected by a variety of factors, including cultural and cognitive biases, such as the well-documented "Matthew effect" [3], [4]. According to this effect, the scientific community pays disproportionate attention to the ideas of already-established scientists [5], [6], making it difficult for competing alternatives to gain attention [7]. These biases slow down the recognition and adoption of important new ideas, resulting in significant time delay in translating new research into new technologies and medical therapies [8], [9]. Yet, examples in which one theory, methodology, or line of inquiry overtakes an established one abound. One such case is the Nobel prize-winning discovery of high-temperature superconductivity in 1986 [10]. This breakthrough challenged the well-established theory of superconductivity [11], which explained how materials enter a superconducting state at low temperatures. Scientists who have been studying superconductors shifted their attention to new materials, which were shown to lose their electrical resistance at much higher temperatures than traditional superconductors, and, therefore, prove to be much more technologically useful. While such shifts are easily recognized in retrospect, many years or decades later, we claim that they are evident in citations patterns almost immediately after the paper describing the breakthrough is published.

Given the importance of timely identification of groundbreaking research, several studies have examined how scientific ideas are adopted by other scientists. Most of these studies analyze citations made by scientific papers, since scientists communicate, position their work, and allocate credit through citations. Using the number of citations, or its distribution, is an accepted way to calculate impact of a paper or a scientist [12]. However, this method is problematic, since it takes years for the citation count to reflect the status of a paper. Mazloumian et al. [13] argued that paradigm shifts occur because an author's groundbreaking paper boosts attention given to his or her other publications. The boost establishes author's "authority," allowing his breakthrough to successfully compete for attention with an established paradigm. They proposed an automatic detection of such boosts as a method for identifying an author's seminal papers.

We view the process by which transformative research is recognized by the scientific community as a competition between paradigms for the attention of the scientific community. A paradigm is a theory of a phenomenon or a research method, *e.g.*, preparation of materials or a new experimental technique. A paradigm is established in one or more papers and supported in subsequent papers. The attention it receives can be measured by the structure of the information cascade the original papers create. The cascade consists of chains of citations that can be traced back to the original papers. We claim that transformative research shifts attention of the scientific community away from the established paradigm and that this is observable as a disruption of the growth of its citations cascade. Disruption occurs when the challenger paradigm can explain new citations received by the established paradigm. Our approach is general and can be applied to other domain, e.g., social media, where ideas compete for attention of information consumers.

Figure 1(a) illustrates our idea. A *seed* (red node) represents a paper establishing some paradigm in a field of research. The paradigm's influence grows over time as new papers cite

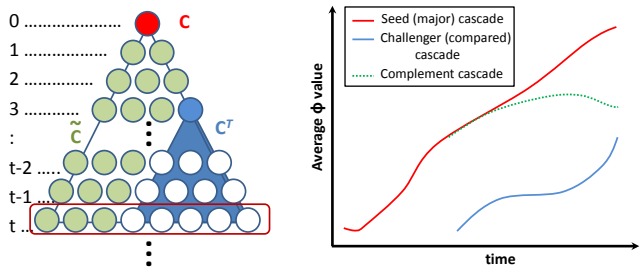


Fig. 1. (a) Information disruption by a challenger in an information cascade. The seed of an established paradigm, marked in red, creates a cascade as it is cited by other papers, while a challenger, marked in blue, disrupts the cascade of the seed. (b) Disruption of the cascade of the seed paradigm (red) by the challenger paradigm (blue) can be visualized as the decline of Φ of the complement cascade (green).

it and are later cited by other papers, creating a *cascade* of citations that can be traced back to the seed. A *challenger* (blue node) is a paper that advocates a new paradigm. It attracts new citations from papers shown as white nodes with blue background, leaving the *complement cascade* (green nodes) containing papers in the cascade of the seed that are not connected to the challenger. When the challenger represents a non-competing idea, though there will be papers that cite both seed and challenger, they will not interfere with the growth of the seed's cascade. In contrast, a transformative challenger will disrupt the growth of the established paradigm. Without considering the challenger, it may appear that the established paradigm continues to prosper, as its cascade continues to grow, but subtracting part of the cascade taken over by the challenger will reveal that the growth of the complement cascade (green nodes) slows. In this case, the community's attention shifts to the challenger paradigm. We propose a method to automatically identify such shifts.

In this paper, we derive an error bound and empirically demonstrate the reliability of our method against sampling fluctuation of the citation network. This is important because complete citations information may not always be available. This property also allows us to scale the method up to large datasets by subsampling.

We illustrate the efficacy of the proposed approach with case studies. Specifically, we selected several highly influential papers from physics and computer science and showed that the proposed method is better able to identify successful challengers than alternative baselines that consider the number of citations received by the paper. Further, we demonstrate that our method identifies challengers that are more relevant to the topic of the seed paper than baselines. Moreover, challenger's success is evident early on, allowing for early detection of transformative research. While the focus of this paper is on scientific publications, the approach can be generalized to other areas where ideas compete to gain attention of information consumers.

II. Approach

We start by defining cascades in citation networks. A citation network is essentially a directed graph $G = (V, E)$ where V is the set of papers and E is the set of edges indicating citations made by papers. A link $(i \leftarrow j) \in E$ denotes that paper j cites paper i, $cite(j)$ denotes the set of all papers that j cites and $cited(i)$ the set of all papers that cite i. V_t is the set of papers published at time t. We assume that if $(i \leftarrow j) \in E$ and $i \in V_t$ and $j \in V_{t'}$ then $t < t'$. That is, no new paper should be cited by an older paper.

Given one or more papers $\mathcal{S} \in G$, a cascade C is a subgraph that contains all citation chains that end at $\mathcal{S}$. The set $\mathcal{S}$ is called the *seed* or *root* of the cascade. The seed indirectly exerts influence on all papers in the cascade, but influence decays with the distance to the seed [14]. For a node j in the cascade, the cascade generating function $\phi(j)$ summarizes the structure of the cascade [15], i.e., all existing citation chains. The cascade generating function quantifies the influence of $\mathcal{S}$ on node j, and is defined recursively by

$$\phi(j) := \begin{cases} 1 & \text{if } j \in S \\ \sum_{i \in cite(j)} \alpha \phi(i) & \text{otherwise,} \end{cases} \tag{1}$$

where α is a constant damping factor. Figure 2 shows an example cascade and the ϕ values for its nodes. For a paper j published after T time steps (*e.g.*, years) from the publication of the seed, $\phi(j)$ can be written as follows:

$$\phi(j) = \sum_{p=0}^{T} a_p \cdot \alpha^p, \tag{2}$$

where the coefficient a_p is the number of distinct paths of length p from one of the seeds to j. The impact of α is that the smaller the value of α, the higher the penalty against long paths. It is also possible to assign a unique α_{ij} for each link but we found that it is simpler to assign a constant 0.5 for all links to control its impact.

A. Cascade Disruption

Consider Figure 1(a). C is the full cascade originated by the seed paper. Let $C^{(\mathcal{T})}$ denote the cascade originating from the challenger $\mathcal{T}$. We define the *complement cascade* $\widetilde{C}$ as the subgraph of C obtained by subtracting $C^{(\mathcal{T})}$ from C, *i.e.*,

$$\widetilde{C} := C - (C \cap C^{(\mathcal{T})}) = C \setminus C^{(\mathcal{T})}.$$

By definition, references of papers in $\widetilde{C}$ can only be traced back to the seed papers but not the challenger. Thus, they represent the influence of $\mathcal{S}$ that cannot be attributed to We note that it is not necessary for the challenger $\mathcal{T}$ to be in C. The blue nodes in Fig. 1(a) are the root node(s) of the intersection of C and $C^{(\mathcal{T})}$. These nodes can be considered as "cross-talk" between the seed and challenger paradigms.

We say that challenger $\mathcal{T}$ disrupts the growth of $\mathcal{S}$ when new papers in the cascade of $\mathcal{S}$ (C) can be explained by the cascade of $\mathcal{T}$ ($C^{(\mathcal{T})}$). This will result in a shrinking complement cascade $\widetilde{C}$. Next, we present a procedure to measure disruption.

Let C_t be the set of papers in cascade C published at time t, *i.e.*, nodes in the bottom red box in Figure 1(a). The average of the cascade function ϕ of papers in C_t is defined by

$$\Phi_t(C) := \frac{1}{|C_t|} \sum_{j \in C_t} \phi(j) = \sum_{p=0}^{t} \overline{a_p} \cdot \alpha^p, \tag{3}$$

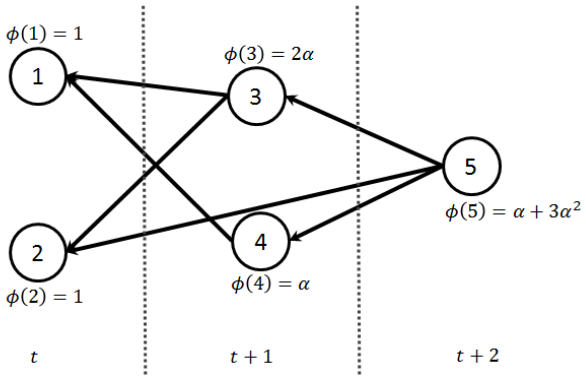


Fig. 2. An example of **Cascade** and their ϕ values.

where $\overline{a_p}$ is the average of the coefficient a_p in Eq. (2) for j in C_t, and $\overline{a_p}$ indicates on average number of distinct citation chains of length p from papers published at time t to the seeds. The variable Φ_t can be interpreted as an indicator of the seed papers' influence at time t.

Figure 1(b) shows the growth of $\Phi_t(C)$ (red curve), $\Phi_t(\widetilde{C})$ (green curve), and the challenger cascade $\Phi_t(C^{(\mathcal{T})})$ (blue curve). The value of Φ_t for both the seed (red) and challenger (blue) papers may both grow rapidly, but the growth of the complement cascade flattens and drops once the challenger successfully shifts the attention of the community. Otherwise, the green curve will continue to grow. In other words, successful information disruption is associated with a declining values $\Phi_t(\widetilde{C})$ of the complement cascade.

We quantify this decline by the *disruption score* $\delta(\tau)$, which is a function of the time interval of τ given the seed and challenger cascades. Let t_0 be the publication time of the challenger paper,

$$\begin{aligned}\delta(\tau) &:= \sum_{t=t_0}^{t_0+\tau} \log \frac{\Phi_t(C)}{\Phi_t(\widetilde{C})} \\ &= \sum_{t=t_0}^{t_0+\tau} \left(\log \Phi_t(C) - \log \Phi_t(\widetilde{C}) \right).\end{aligned}$$

The disruption score can be visualized as the area between the red and green curves in Figure 1(b) from t_0 to $t_0 + \tau$. The disruption score allows us to identify and measure the impact of the challenger paper.

When comparing candidate challengers published too long apart over time, the cascade of the seed paper may be so different that might give unfair advantages to old challengers. For example, the cascade of a seed paper published in 1950's may grow many-fold from 60's to 90's. For a new paper to disrupt the same proportion of the cascade as an old paper may require a much larger number of citations. The disruption score is immune from this problem because ϕ will be smaller after 30 years as citation paths to the seed stretch. More importantly, we consider the average, not sum. Also, the number of publications and thus citations to new papers grow faster in recent years and may compensate for the difference in cascade size.

B. Computing Cascade Disruption

To obtain the disruption score, we need to compute ϕ of the nodes in the cascade. A citation network is a directed acyclic graph if cycles are considered as errors. From Eq. (1), traversing the citation network in a topological order [16] and updating ϕ values along the way will guarantee that no backtracking is necessary to compute all ϕ values for all nodes. Therefore, we can apply topological sorting to compute ϕ and obtain the disruption scores. The time complexity of topological sorting is O($|V_C|$+$|E_C|$), which is linear to the sum of the number of nodes and edges in cascade C.

Cascade generating function ϕ can measure information cascades not only in citations networks, but also in other domains, such as information diffusion in social media or influence in social networks. The method for measuring disruptions of cascade growth should, therefore, carry over to these domains as well. This could lead to numerous other applications, such as comparing competing memes that are spreading in social media to determine which one is attracting more attention, or which person is becoming more influential.

III. Evaluation

According to the classical test theory, a quantitative measure must be both valid and reliable. The notion is closely related to bias and variance in statistical data mining and pattern recognition [17]. As a quantification measure of transformative research, the disruption score must be *valid*, in the sense that truly transformative research will be scored higher than others, and *reliable*, in the sense that the score is robust against incomplete subsampling of citation network data. In addition, computation of the score must scale up to large citation network data. In this section, we evaluate the validity, reliability, and scalability of the disruption score as a detector of transformative research.

A. Data

We use two large citation network datasets in the empirical evaluation. One of them is the dataset of the journals published by the American Physical Society (APS) [18],which consists of articles published from 1893 to 2009. The APS dataset contains important physics papers that announced a new discovery or a new technique, many of which were recognized by the Royal Swedish Academy of Sciences with a Nobel prize, the highest honor in physics. APS is perfect for our study because it contains many examples of successful transformative research and recognized paradigm shifts and makes them available for analysis.

The other dataset is the DBLP-Citation-network V5 (DBLP) available at Arnetminer.org [19], [20], [21], [22], which consists of two major computer science bibliographic datasets, DBLP and ACM, covering publications from 1936 to 2011. The DBLP dataset contains some of the important papers in computer science that describe widely used techniques and algorithms. Table I summarizes the statistics of these datasets. The difference of the network structure and the scale of these two datasets reflects the difference in citation culture between these two disciplines of science.

To reduce noise, we pruned low-citation publications as a pre-processing step. Papers must be cited more than 10 times

TABLE I. STATISTICS OF THE TEST DATA

dataset	# paper	# citations	avg. degree
APS	449,667	4,710,548	20.91
pruned	115,753	1,153,967	19.02
DBLP	1,572,278	2,083,947	2.65
pruned	82,762	414,776	10.46

in APS and 5 in DBLP to be included in our evaluation. We considered a citation to a more recent paper as an error and removed 284 from APS and 20,418 from DBLP, respectively. In addition, we excluded 2,555 review articles from the APS dataset that were published in *Reviews of Modern Physics*, since their citation patterns are different from regular research papers [23]. Review papers never start a new paradigm or become a challenger by definition and thus are not in the scope of our search.

B. Validity

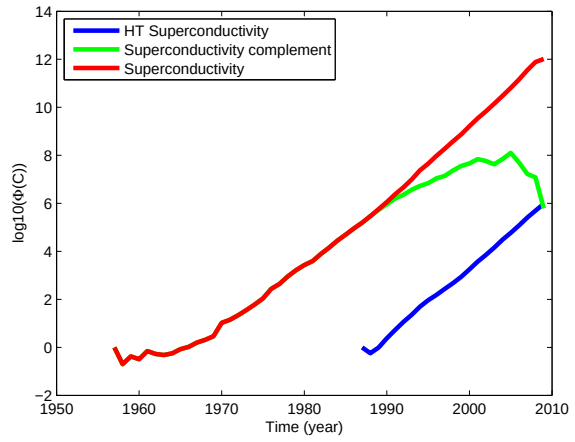
The disruption score is a valid indicator of paradigm shift if the score distinguishes truly transformative research papers from the rest with high sensitivity and specificity. However, unlike well-defined data mining problems, it is difficult to create a large gold standard of truly transformative research to quantitatively assess the validity of the proposed method. Therefore, we focus on a few well-known cases of transformative research to evaluate our method's validity. Section IV reports detailed results of applying our method to APS and DBLP datasets.

Consider superconductivity. The 1957 theory of superconductivity by Bardeen, Cooper, and Schriffer (BCS) [11], [24] was a dominant paradigm in this field until the discovery of high-temperature superconductivity [10] (HTS) in 1986, an indisputable transformative research accomplishment for which the authors were awarded the Nobel Prize in Physics the next year.

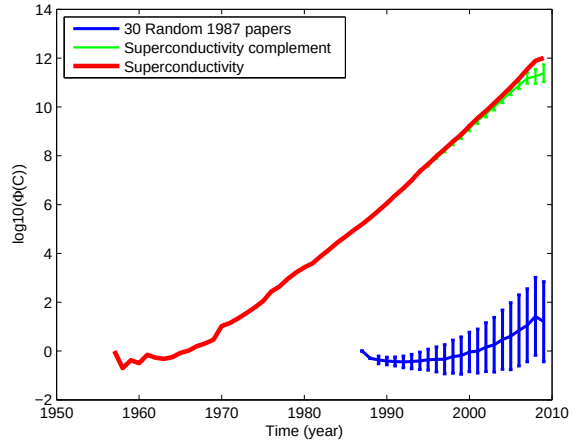
Figure 3(c) shows evolution of the cascade size, *i.e.*, the number of papers in the cascade, of the BCS cascade, and the HTS cascade rooted at three pioneering APS papers in this field [25], [26], [27]. One may expect that discovery of HTS would slow down the growth of the BCS cascade, but Figure 3(c) shows otherwise. The cascade size, in terms of the cumulative number of papers in the cascade each year, continues to grow, though at a slower pace than HTS. HTS might surpass BCS soon, but the impact of paradigm shift is hardly observable 20 years later if we use the cascade size as an indicator.

Figure 3(a) compares the growth of the logarithms of $\Phi_t(C^{(bcs)})$ (red), $\Phi_t(C^{(hts)})$ (blue) and $\Phi_t(\widetilde{C})$ (green) as computed from the APS dataset. We see a pattern identical to the one shown in Figure 1(b), a vivid demonstration of cascade disruption and paradigm shift. Moreover, the disruption starts immediately after the publication of HTS.

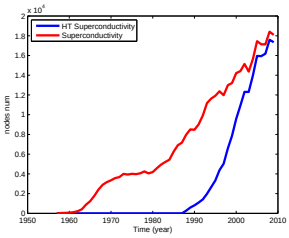
To test the specificity of cascade interruption, we randomly selected 30 papers published in 1987, the same year as HTS seeds, from the APS dataset as negative controls and plotted the growth of their cascades as shown in Figure 3(b), where the blue curve shows the means and standard deviations of the average cascades of these 30 challengers and the green curve shows those for their complement cascades. The curves show



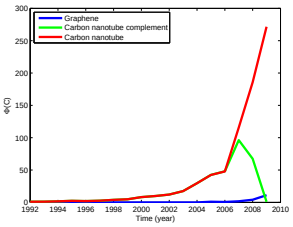
(a)



(b)



(c)



(d)

Fig. 3. **Cascade disruption as an indicator of paradigm shift**. The growth of the logarithm of average cascade function values per year for (a) superconductivity (BCS) [11], [24] and high-temperature superconductivity (HTS)[25], [26], [27] and (b) 30 control cases published in 1987 show no sign of cascade disruption against BCS. (c) The growth of the size of cascades shows no sign of disruption. (d) Another example of paradigm shift is conventional carbon nanotube [28] versus graphene [29], [30]. Unlike superconductivity, the cascades are not as large because they were published in recent years. Therefore, no logarithm of Φ is taken here.

that though the growth of their cascades varies widely, the complements of the BCS cascade are hardly disrupted, unlike the HTS papers.

Another example of transformative research is the development of graphene in 2004 [29], [30], which was considered a breakthrough both for the materials fabrication technology,

focused on carbon nanotubes [28] and as a system for studying properties of 2-dimensional electron systems. The developers of graphene were awarded Nobel Prize in Physics in 2010. Figure 3(d) shows the cascade growth and disruption in this case. Again, the disruption is observable starting in 2006, right after their publication. This is as fast as possibly detectable because we removed all citations between papers published in the same year. The disruption then drops sharply in 2007, three years before their Nobel Prize award, even though the growth of the average cascade of the graphene papers is flat.

C. Reliability

Existing datasets of citation networks are inevitably incomplete and only contain a subset of all related papers and citations. It is important that the proposed disruption score produces consistent results given different subsamples of citation network data.

Here we show that it is possible to derive a theoretic error bound of the disruption score given a subsample of citation network data, compared to the score obtained from the complete citation network. We observed empirically in our preliminary study that if the average cascade function values $\Phi_t(C) > \Phi_t(\widetilde{C})$, the relation will maintain when they are estimated from a subsampled cascade, *i.e.*, $\Phi_t(C') > \Phi_t(\widetilde{C'})$. In other words, if we observe cascade disruption in a subsampled cascade, then it is almost certain that cascade disruption will also present in a complete cascade.

To see why this is the case, let C' be the subsampled cascade from the complete cascade C with a constant node sampling ratio ρ. The citation links in C adjacent to nodes not in C' are removed from C'. Since Φ essentially is the true mean of the cascade function values ϕ given a complete cascade C, if ϕ of the nodes in the subsampled cascade C' are identical to their ϕ values in the the complete cascade C, then according to Hoeffding's inequality, which states that the probability that the difference of sample mean and true mean is large is less than a formula that is roughly proportional to the exponential of the inverse of the sample size, we can show that $\Phi_t(C') \approx \Phi_t(C)$ and this is similarly the case for the complement cascade $\widetilde{C}$ and its subsample $\widetilde{C'}$ and hence the inequality relation will maintain.

However, the new ϕ of the nodes in the subsampled cascade C' will be different, because citations to the removed nodes are absent. Also, ϕ of those being cited will be smaller due to the removal of the unselected nodes. Therefore,

$$\begin{aligned}\Delta\phi_C(j) &= \phi_C(j) - \phi_{C'}(j) \\ &= \alpha\left(\sum_{i\in C}\phi_C(i)I(i \in cite(j) \& i \notin C') \right. \\ &\left. + \sum_{i\in C}\Delta\phi_C(i)I(i \in cite(j) \& i \in C')\right),\end{aligned}$$

and its expectation will be

$$\begin{aligned}&\mathbb{E}[\Delta\phi_C(j)] \approx \\ &\alpha\mathbb{E}(|cite(j)|)|C|\big((1-\rho)\mathbb{E}[\phi_C(i)] + \rho\mathbb{E}[\Delta\phi_C(i)]\big),\end{aligned} \tag{4}$$

(a) (b)

Fig. 4. Heatmaps of correlation between trials of 5-fold cross-validation for (a) "Fast Algorithms for Mining Association Rules in Large Databases (1994)" and (b) "Induction of Decision Trees (1986)". "Full" is the result for the complete data, "Ex.CV fold i" is the result of the i-th cross validation trial.

where $|C|$ is the number of nodes in cascade C. This applies to the complement cascade $\widetilde{C}$ and its subsample $\widetilde{C'}$ as well. Since Φ is the expectation of ϕ for papers published at the same time, from the Hoeffding's inequality and Eq. (4), we can conclude that with a high probability proportional (roughly speaking) to the sampling size of the subsampled cascade, the difference

$$|\Delta\Phi_t(C) - \Delta\Phi_t(\widetilde{C})| = |\Phi_t(C') - \Phi_t(\widetilde{C'}) - (\Phi_t(C) - \Phi_t(\widetilde{C}))|$$

will be very small. The following theorem establishes a bound for the sampling error of the average cascades.

Theorem 1. *For any strictly positive constant ε, with probability greater than*

$$1 - 2e^{-2\varepsilon^2|C'_t|}\frac{1}{4\varepsilon|C'_t|}\sqrt{\frac{\pi}{2|C'_t|}} - 2e^{-2\varepsilon^2|\widetilde{C'_t}|}\frac{1}{4\varepsilon|\widetilde{C'_t}|}\sqrt{\frac{\pi}{2|\widetilde{C'_t}|}},$$

$$\begin{aligned}&\textit{if } \big(\Phi_t(C') - \Phi_t(\widetilde{C'})\big) > 0, \textit{then} \\ &\qquad \exists S > 0, \big(\Phi_t(C) - \Phi_t(\widetilde{C})\big) \in (S+\varepsilon, S-\varepsilon).\end{aligned}$$

Proof: Appendix provides a sketch of proof by the Hoeffding's inequality. ■

We also empirically tested the reliability of the disruption score with subsampling. We chose the top highly cited papers in DBLP and ranked the papers in their cascades according to their disruption scores. Next, we assessed the reliability of our method by a 5-fold cross validation sampling test, where we divided all papers in the dataset into five subsets and used four of them to assign the ranks. Then we used the Spearman's rank correlation coefficient to measure the similarity of the ordering of the top 1000 articles in the five trials. The similarity tests show that using 80% of the data yields similar disruption scores and similar rankings. Figure 4 shows the heatmaps of correlations for two well-known papers in the data mining community as the seeds. We also observed that the differences between the disruption scores of the top 5 challengers computed from the cross-validation subsamples and from the complete dataset are small and with negligible variance (data not shown). We set $\tau = 4$ years when computing the disruption score in all trials. Using other highly-cited papers in DBLP gives similar results.

D. Scalability

We already showed that the algorithm to compute the disruption score is linear in the size of the citation network. As the number of publications grew geometrically in recent years, and to apply the algorithm to even larger networks of social media, the scalability of the algorithm has to improve further. One of the options is to explore subsampling. Theorem 1 and the empirical results show that the disruption score can be estimated reliably from a subsampled citation network. This useful property allows us to further accelerate computation. With a suitable sampling, computing the disruption scores can be more efficient in both computation time and memory space. According to our execution time statistics with different ratios of node sampling from APS, the time drops nearly exponentially because the number of citations decreases exponentially as the number of nodes decreases linearly: e.g., sampling 80% of papers can save 55% of the time.

When the task is to rank a large number of candidate challengers by their disruption scores, it is possible to avoid exhaustive pairwise comparison by reusing intermediate results. Suppose we would like to rank 100 candidate challengers by their disruption scores. A brute-force approach is to compute the complement cascades for each of the candidates. By sorting these candidates in their topological order in the citation network, the ϕ values computed for the upstream candidates can be reused for the downstream candidates and significantly reduce the computational costs.

IV. Results

In previous section, we report an evaluation of our method by testing if the disruption scores are high for known examples of transformative research when they are scored against the representative papers of the paradigms that were disrupted. In this section, we report a further test, where a highly cited paper is chosen and the goal is to use our method to rank all the papers in its cascade by their disruption score and see whether the highest scoring paper represents the best transformative research, under the condition that the system is blind about which papers are transformative. We note that in this case, it is possible that no challenger is sufficiently transformative against selected high cited papers but the highest scoring ones may still hint us about which papers are emerging. Again, since it is difficult to create a large set of the "ground truth" of transformative research, we will not to provide a quantitative evaluation, such as measuring error rates or area-under-curve, but will demonstrate through several case studies that the proposed method is able to identify examples of transformative research. We use the APS and DBLP datasets described in Section III-A to identify examples of transformative research in physics and computer science, respectively. We compare our method to two *baselines* and show that our method identifies more relevant challengers than both baselines.

A. Baselines

We compare our method to *citations baseline* that orders the papers within a cascade by their popularity, *i.e.*, the number of citations they received. We also compare our method to *cover ratio baseline*: given a challenger $\mathcal{T}$, cover ratio of $\mathcal{T}$ is defined as,

TABLE II. Top ten challengers to the 1957 "Theory of Superconductivity" identified by (a) proposed method and (b) baseline.

Year	Citations	CoverRatio	Title
(a) our method: sorted by disruption score			
1958	14	0.97	Meissner Effect
1958	307	0.98	Random-Phase Approximation in the Theory ...
1959	40	0.92	Evidence for Anisotropy of the Superconducting ...
1989	574	0.25	Phenomenology of the normal ...
1987	368	0.35	Antiferromagnetism in La_2CuO_{4-y}
1987	281	0.35	Two-dimensional antiferromagnetic quantum ...
1988	149	0.31	$Ba_2YCu_3O_{7-\sigma}$: Electrodynamics ...
1990	156	0.19	High-resolution angle-resolved photoemission ...
1988	399	0.28	Low-temperature behavior of two-dimensional ...
1995	95	0.06	Momentum Dependence of the Superconducting ...
(b) baseline: sorted by cover ratio			
1958	307	0.98	Random-Phase Approximation in the Theory ...
1958	14	0.97	Meissner Effect
1958	63	0.96	Coherent Excited States in the Theory of ...
1958	93	0.96	Paramagnetic Susceptibility in Superconductors
1958	14	0.96	Meissner Effect and Gauge Invariance
1960	246	0.96	Quasi-Particles and Gauge Invariance in the ...
1959	36	0.94	Impurity Scattering in Superconductors
1959	37	0.94	Collective Excitations in the Theory of Supercond...
1958	119	0.93	Possible Analogy between the Excitation Spectra ...
1960	32	0.93	Magnetic Moment of Transition Metal Atoms in ...
(c) baseline: sorted by citations			
1981	3191	0.48	Self-interaction correction to density-functional ...
1996	3088	0.05	Generalized Gradient Approximation Made Simple
1980	2651	0.51	Ground State of the Electron Gas by a Stochastic ...
1976	2569	0.55	Special points for Brillouin-zone integrations
1996	2387	0.02	Efficient iterative schemes for ab initio total-energ...
1990	1951	0.12	Soft self-consistent pseudopotentials in a generaliz...
1991	1950	0.13	Efficient pseudopotentials for plane-wave calculations
1975	1597	0.58	Linear methods in band theory
1992	1567	0.09	Atoms, molecules, solids, and surfaces: Applications ...
1992	1445	0.09	Accurate and simple analytic representation of the ...

$$CoverRatio(\mathcal{T}) := \frac{|(C \cap C^{(\mathcal{T})})|}{|C|} \tag{5}$$

i.e., the ratio of nodes in the cascade that can be traced to the challenger $\mathcal{T}$.

B. Physics

We chose several papers with the most citations in our dataset, which came from different subfields of physics. We identified the most disruptive challengers of these papers and carried out quantitative analysis of their topics.

a) Case Study 1: In 1957 Bardeen, Cooper and Schrieffer published a seminal paper titled "Theory of Superconductivity" which explained the mechanism by which some metals became perfect electrical conductors (*i.e.*, they lost their electrical resistance) at low temperatures. The authors were awarded a Nobel prize for this discovery in 1972. This paper is one of the ten most cited papers in the APS dataset. Table II lists the ten top-ranked challengers identified by the method proposed in this paper and the baseline. The disruption score of challengers was computed for a ten-year period ($\tau = 10$). Compared to the citations baseline, both our method and the cover ratio baseline identifies papers that are relevant to the topic of superconductivity. All ten of the top challengers identified by baseline are papers dealing with calculations of electronic structure of materials, and include other most-cited papers in the APS dataset. While this is a very important topic, it is only peripherally related to superconductivity, in as much as this phenomenon is a result of correlated electron pairs.

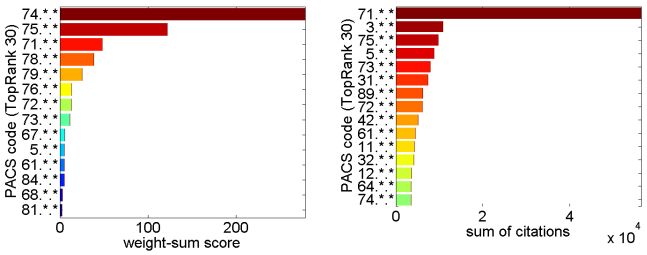


Fig. 5. Distribution of PACS numbers of challenger papers identified by (a) our method and (b) baseline

While top-rated challengers discovered by the cover ratio baseline are on the topic of superconductivity, only the proposed method discovered papers on high temperature superconductivity (HTS). The discovery of HTS was an important development in the study of superconductivity, recognized with a Nobel prize in 1987. Although the original paper announcing the discovery is not in our dataset, presence of several other papers on HTS among the top challengers demonstrates the efficacy of our method to identify disruptive papers. These challengers include "Antiferromagnetism in La_2CuO_{4-y}", "Two-dimensional antiferromagnetic quantum spin-fluid state in La_2CuO_4", "$Ba_2YCu_3O_7$: Electrodynamics of Crystals with High Reflectivity" and "Momentum Dependence of the Superconducting $Sr_2CaCu_2O_8$".

We verify that our method identifies more relevant challengers through the analysis of their PACS numbers. The Physics and Astronomy Classification Scheme (PACS) was introduced in 1975 to allow authors to identify the field and subfields of their papers. Figure 5 shows the frequency distribution of PACS categories of the 30 top-ranked challengers (with PACS numbers) identified by our method and the citations baseline, weighted by the score assigned to the paper by the method. We aggregated the numbers by their top-level category. Our method finds many more papers about "Superconductivity" (category 74) and "Magnetic properties and materials" (75) than the citations baseline, which found papers about "Electronic structure of bulk materials" (71), "Quantum mechanics" (3) and "Statistical physics" (5), while "Superconductivity" (topic 74) is 15th most frequent topic among these challengers. In fact, topics (71, 75, 5) are the most common PACS numbers in the entire dataset, suggesting that citations baseline picks out globally popular papers, even though it considers only the papers in the cascade created by the seed node. None of the top-30 challengers identified by the cover ratio baseline had PACS numbers, since they were published before the classification system was introduced.

b) Case Study 2: We used our method to rank challengers of the most cited paper in particle physics[1]. This is the 1967 paper by Steven Weinberg titled "A Model of Leptons." This seminal work unified weak and electromagnetic interactions within a single theory of electroweak interactions. It won its authors a Nobel prize in 1979.

Table III lists the ten challengers to this paper with highest disruption score, which was computed for $\tau = 10$ years. The

[1]http://www.slac.stanford.edu/spires/topcites/2010/alltime.shtml

TABLE III. TOP TEN CHALLENGERS TO THE 1967 "A MODEL OF LEPTONS" BY STEVEN WEINBERG IDENTIFIED BY (A) PROPOSED METHOD AND (B) BASELINE METHOD.

Year	Citations	CoverRatio	Title
(a) our method: sorted by disruption score			
1974	190	0.88	Asymptotically free gauge theories. II
1974	123	0.88	Electroproduction scaling in an asymptotically ...
1974	309	0.78	Hierarchy of Interactions in Unified Gauge ...
1974	696	0.89	Confinement of quarks
1973	162	0.88	New Approach to the Renormalization Group
1972	46	0.89	Spontaneous Breakdown and Hadronic Symmetries
1973	44	0.88	Unified Gauge Theories of Hadrons and Leptons
1972	208	0.90	Effects of a Neutral Intermediate Boson in ...
1974	361	0.80	Experimental Observation of a Heavy Particle J
1973	59	0.79	Current Algebra and Gauge Theories. I
(b) baseline: sorted by cover ratio			
1971	236	1.00	Physical Processes in a Convergent Theory of ...
1970	35	1.00	Spontaneous Breakdown of Weak and Electromagn...
1972	95	0.99	Renormalizable Massive Vector-Meson Theory-...
1972	71	0.99	Short-Distance Behavior of Quantum Electrody...
1973	742	0.99	Radiative Corrections as the Origin of ...
1972	157	0.99	Spontaneously Broken Gauge Symmetries. I. ...
1972	109	0.99	Spontaneously Broken Gauge Symmetries. II. ...
1972	68	0.99	Approximate Symmetries and Pseudo-Goldstone ...
1972	94	0.99	Spontaneously Broken Gauge Symmetries. III. ...
1972	58	0.98	Mixing Angle in Renormalizable Theories of ...
(c) baseline: sorted by citations			
1981	3191	0.61	Self-interaction correction to density-functional ...
1996	3088	0.07	Generalized Gradient Approximation Made Simple
1996	2387	0.03	Efficient iterative schemes for ab initio total-...
1990	1951	0.15	Soft self-consistent pseudopotentials in a ...
1991	1950	0.17	Efficient pseudopotentials for plane-wave ...
1992	1567	0.12	Atoms, molecules, solids, and surfaces: ...
1992	1445	0.11	Accurate and simple analytic representation ...
1994	1430	0.04	Projector augmented-wave method
1999	1424	0.01	From ultrasoft pseudopotentials to the projector ...
1993	1345	0.06	Ab initio molecular dynamics for liquid metals

first and second challengers are papers by David Gross and Frank Wilczek, and David Politzer respectively. These three physicists shared a Nobel prize in 2004 for elucidating the theory of strong interactions, which along with gravity, electromagnetic, and weak interactions forms the four fundamental forces of nature. Though they did their works independently and their papers received a nod from the Nobel committee 30 years after their publication, our method identifies both of them and correctly recognizes their contributions already using the data less than ten years after their publication.

The top-ten challengers include four more papers by Steven Weinberg (3, 5, 8, and 10), and papers by Nobel laureates Kenneth G. Wilson (paper 4 "Confinement of quarks"), and Samuel C. Ting (paper 9 "Experimental Observation of a Heavy Particle J"). Though these papers are on slightly different topics than the seed, they are all important works within the particle physics community, demonstrating that our method is able to capture how the community shifts its attention between different topics.

The cover ratio baseline, on the other hand, identifies papers on several topics of particle physics, most notably gauge symmetry breaking. This is an important research area in theoretical particle physics, but one that generalizes to all forces. In contrast, our method found more challengers relevant to the topic of electroweak interactions.

Of the top ten challengers identified by citations baseline, seven are the same as the baseline challengers of the first case study. These are papers in a popular topic of using density functional theory, or its variants, for electronic structure calculations, and are not relevant to the topic of high energy

TABLE IV. TOP TEN CHALLENGERS (PUBLISHED AFTER 1994) TO THE 1982 "TWO-DIMENSIONAL MAGNETOTRANSPORT IN THE EXTREME QUANTUM LIMIT" IDENTIFIED BY (A) PROPOSED METHOD AND (B) BASELINE.

Year	Citations	CoverRatio	Title
(a) our method: sorted by disruption score			
1995	246	0.13	Spontaneous interlayer coherence in double-...
2005	179	0.01	Unconventional Integer Quantum Hall Effect ...
2005	65	<0.01	Electric Field Modulation of Galvanomagnetic ...
2005	178	<0.01	Quantum Spin Hall Effect in Graphene
1995	199	0.11	Optically Pumped NMR Evidence for Finite-Size ...
2005	42	<0.01	Coulomb interactions and ferromagnetism in pure ...
2005	21	<0.01	Disorder and interaction effects in two-dimensio...
2005	14	<0.01	Coexistence of sharp quasiparticle dispersions ...
2005	45	<0.01	Local defects and ferromagnetism in graphene ...
2005	121	<0.01	Z_2 Topological Order and the Quantum Spin Hall ...
(b) baseline: sorted by cover ratio			
1995	85	0.17	Updated analysis of πN elastic scattering data ...
1996	857	0.17	Review of Particle Physics
1995	50	0.17	πN-ηN and πN-ηN partial-wave T matrices in ...
1995	51	0.17	Baryon current matrix elements in a light-front ...
1995	22	0.17	Kinematic evidence for top quark pair production ...
1995	18	0.17	Search for High Mass Top Quark Production in pp...
1995	269	0.17	Observation of the Top Quark
1995	337	0.17	Observation of Top Quark Production in p...
1995	11	0.17	πN S-wave scattering length in a three-coupled-cha...
1995	72	0.15	Static Response and Local Field Factor of the Elec...
(c) baseline: sorted by citations			
1996	3088	0.12	Generalized Gradient Approximation Made Simple
1996	2387	0.05	Efficient iterative schemes for ab initio total-en...
1999	1424	0.02	From ultrasoft pseudopotentials to the projector au...
1998	1003	0.10	Quantum computation with quantum dots
1996	857	0.17	Review of Particle Physics
1998	845	0.04	Entanglement of Formation of an Arbitrary State of Two...
1996	795	0.09	Mixed-state entanglement and quantum error correction
1998	748	0.05	Evidence for Oscillation of Atmospheric Neutrinos
1998	737	0.08	Cold Bosonic Atoms in Optical Lattices
1995	664	0.12	Double Exchange Alone Does Not Explain the Resistiv...

physics. This case study further highlights the ability of our method to identify important and relevant challengers.

c) Case Study 3: Our final study considers the fractional quantum Hall effect, a phenomenon in which the conductance of 2-dimensional electrons is quantized at certain levels. This effect was first reported in a 1982 paper titled "Two-Dimensional Magnetotransport in the Extreme Quantum Limit." The discovery and explanation of this effect was recognized with a Nobel prize in 1998. Table IV shows the top ten challengers identified by our method and baseline. Since the APS dataset ends in 2009, the disruption score for 2005 papers were computed for the four year period. To mitigate the bias that incomplete data introduces, we show only the challengers published after 1994.

Several of the challengers with highest disruption score are about graphene, a one-atom-thick layer of graphite, whose discovery has facilitated new investigations of the properties of matter and electrons confined to 2-dimensional surfaces, and resulted in a Nobel prize in 2010. In comparison, both baseline methods identify irrelevant challengers, including those dealing with the top quark (papers (b)6–8), calculations of electronic structure of bulk materials (papers (c)1–3), quantum computing (papers (c)4, 7) and high energy physics ((b)2, (c)5, 8, 9).

d) Quantitative Analysis: We validate quantitatively that the proposed method identifies more relevant challengers than baseline by performing PACS number analysis of the challengers for the ten most-cited papers in the APS dataset. We compared the PACS number distribution of the 30 top-ranked challengers identified by each method with the distribution of PACS numbers of all papers in the APS dataset (with PACS numbers). The mean correlation of the distributions of PACS numbers of challengers of the 10 most-cited APS papers identified by our method with the global PACS number distribution is 0.4611 ± 0.0048. The mean correlation of PACS number distribution of top challengers of the most-cited papers found by the citations baseline with the global PACS number distribution is 0.5800 ± 0.0033. Higher correlation of the citations baseline indicates that it tends to identify challengers on globally popular topics, compared to the proposed method, which tends to identify challengers that are topically relevant to the seed.

TABLE V. TOP CHALLENGERS TO THE 1986 "INDUCTION OF DECISION TREES" PAPER IDENTIFIED BY (A) PROPOSED METHOD AND (B) BASELINE METHOD.

Year	Citations	CoverRatio	Title
(a) our method: sorted by disruption score			
1995	189	0.33	Discovery of Multiple-Level Association Rules...
1995	227	0.35	An Effective Hash Based Algorithm for Mining ...
1996	211	0.29	Sampling Large Databases for Association Rules.
1995	254	0.32	An Efficient Algorithm for Mining Association ...
1997	191	0.15	Beyond Market Baskets: Generalizing Association ...
(b) baseline: sorted by cover ratio			
1992	53	0.56	Querying in Highly Mobile Distributed Environments.
1993	33	0.52	Relevance Feedback and Inference Networks.
1992	64	0.51	An Interval Classifier for Database Mining Appli...
1993	143	0.50	Database Mining: A Performance Perspective.
1993	1372	0.47	Mining Association Rules between Sets of Items ...
(c) baseline: sorted by citations			
1994	1592	0.45	Fast Algorithms for Mining Association Rules in ...
1993	1372	0.47	Mining Association Rules between Sets of Items ...
2000	647	0.11	Directed diffusion: a scalable and robust communic...
2002	602	0.01	Wireless sensor networks: a survey.
2000	523	0.02	Content-Based Image Retrieval at the End of the Ear...

C. Computer Science

We report results of three case studies of high interest to the data mining community, using the most highly cited papers in the DBLP dataset as seeds. Due to the fast pace of computer science research, we set $\tau = 4$ years to compute the disruption score $\delta(\tau)$.

e) Case Study 1: Ross Quinlan's 1986 paper on ID3 is one of the most influential papers in computer science that laid the foundation of the field of classifier learning. One may expect that papers about new algorithms of classifier learning are top challengers that transform the field, but surprisingly the results given in Table V shows that it is papers about association rule mining. Research in association rule mining led to a whole new field of data mining, while in comparison, new research in classifier learning is still within the realm laid out by Quinlan's ID3. In this sense, our result is more reasonable. The top challenger is perhaps the most related to ID3 among papers on association rule mining because a decision tree can be considered as a set of multiple-level rules. Our results are also more reasonable than those found by both baselines. Though two association rule mining papers appear in the top-5 lists found by the challengers, the remaining papers are irrelevant to decision tree learning. We note that the disruption scores of all of the identified papers are not very high and the top ones are close, suggesting that no single truly transformative research has yet challenged classic decision tree learning.

f) Case Study 2: Next, we asked what challenges association rule mining. Our seed selection is Agrawal and

TABLE VI. TOP CHALLENGERS TO THE 1994 "FAST ALGORITHMS FOR MINING ASSOCIATION RULES IN LARGE DATABASES" PAPER IDENTIFIED BY (A) PROPOSED METHOD AND (B) BASELINE METHOD.

Year	Citations	CoverRatio	Title
(a) our method: sorted by disruption score			
1995	227	0.78	An Effective Hash Based Algorithm for ...
1995	189	0.74	Discovery of Multiple-Level Association ...
1996	211	0.65	Sampling Large Databases for Association ...
1995	254	0.71	An Efficient Algorithm for Mining Associa...
1998	170	0.25	Exploratory Mining and Pruning Optimizatio...
(b) baseline: sorted by cover ratio			
1995	227	0.78	An Effective Hash Based Algorithm for Mining ...
1995	189	0.74	Discovery of Multiple-Level Association Rules ...
1995	254	0.71	An Efficient Algorithm for Mining Association ...
1996	211	0.65	Sampling Large Databases for Association Rules.
1997	252	0.58	Dynamic Itemset Counting and Implication Rules for ...
(c) baseline: sorted by citations			
2000	523	0.05	Content-Based Image Retrieval at the End of the ...
2000	492	0.07	Mining Frequent Patterns without Candidate Gener...
2002	350	0.05	Optimizing search engines using clickthrough data.
2002	338	0.08	Models and Issues in Data Stream Systems.
2001	328	0.05	Item-based collaborative filtering recommendation ...

Srikant's 1994 seminal paper, which is the third most-cited paper in DBLP. The results, shown in Table VI, suggest that it remains dominant in data mining, as top five challengers are all follow-up papers with relatively low disruption scores (data not shown). Here, cover ratio baseline identifies similar challengers as those found by our method. The citations baseline selects mostly irrelevant papers. It is commonly agreed that the hash-based algorithm is truly a breakthrough for association rule mining and its disruption score correctly reflects that.

V. RELATED WORK

Much of bibliometric analysis uses citations count to measure a paper's quality or scientist's productivity [12]. Beyond simple citations count, researchers have explored methods that analyze the structure of citation networks to identify important papers [31], [32] or predict which papers will be important in the future [33].

Few works have explicitly studied transformative research or develop methods to automatically identify such research. Mazloumian et al.[13] characterized how a publication of a landmark paper increases attention paid to author's other papers, leading to a paradigm shift, which may eventually be recognized with a Nobel prize. Chen [34] described the use of a dynamic co-citation network to reveal "intellectual turning point" papers. Our approach differs from related work in that, first, we explicitly target papers that disrupt established works, and second, we consider cascades, which take chains of citations into account. Next, it is well-known that citation counts decay over time even for a highly influential work [35]. Therefore, it is important to consider its continuing influence of cascades, which provide indirect exposure to the work. Ghosh and Lerman [15] developed a function to quantify the structure of a growing cascade of information spreading in social media, which we use to measure the size of evolving cascades. In this paper we propose an efficient method that use this function to identify transformative scientific research.

How information spreads in a network of information ecosystems like social media and scientific publications has been heavily studied. Various models are available to explain and predict information diffusion [15], [36], [37], [38]. Widely spread information may be disrupted by the presence of another piece of information that competes to gain attention from information consumers [39]. In scientific publications, information diffusion is usually measured by counting citations, and citations-based measures, such as the h-index [12], are widely used to evaluate the productivity of scientists. Disruptions of citation cascade growth of well-established, field-defining papers usually represent an event of "paradigm shift," "breakthrough," emergence of a "disruptive idea," and a successful "transformative research." Similarly, in social media, the flow of a dominant topic may be disrupted by a challenger, which will gain attention from crosstalk information consumers, who have been following the dominant topic but now switch their attention. A challenger successfully disrupts the dominant topic when the challenger substitutes the attention of a sufficient number of crosstalk information consumers.

VI. CONCLUSIONS AND FUTURE WORK

Transformative research shifts attention of the scientific community from the established paradigms that represent theories and methods accepted and practiced by the community. The degree to which the paradigm is accepted by the community is reflected in the citations received by papers that first describe it, and citations received by these papers, and so on. By looking at the structure of the citations cascade, we can determine when a new paradigm attracts attention of the scientific community. This happens when citations received by papers advancing the new paradigm can explain most of the new citations received by the old paradigm. These shifts of attention are evident soon after the challengers' publication, enabling early detection of transformative research. We have proposed a method to identify transformative challengers, i.e., scientific papers that shift attention of the community, by measuring how much they disrupt the growth of citation cascades of papers representing the established paradigm. When applied to citations networks of physics and computer science papers, our method correctly identified several examples of transformative research.

More work needs to be done to elucidate the processes that lead to shifts of attention. We need to identify seeds which simply do not have any significant challengers. Also, we would like to develop scalable methods that take into a account a set of seeds and a set of challengers. Another interesting direction is to develop methods to identify which established idea a given paper disrupts. We believe that identifying transformative research by analyzing citations cascades will prove to be a productive line of inquiry.

REFERENCES

[1] N. S. B. (U.S.), *Enhancing support of transformative research at the National Science Foundation [electronic resource]*. National Science Foundation, Arlington, VA :, 2007.

[2] T. S. Kuhn, *The Structure of Scientific Revolutions: 50th Anniversary Edition*, fourth edition ed. University Of Chicago Press, 2012.

[3] R. K. Merton, "The Matthew Effect in Science," *Science*, vol. 159, no. 3810, pp. 56–63, Jan. 1968. [Online]. Available: http://dx.doi.org/10.1126/science.159.3810.56

[4] ——, "The matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property," *Isis*, vol. 79, no. 4, pp. 606–623, 1988. [Online]. Available: http://dx.doi.org/10.2307/234750

[5] P. D. Allison, "Inequality and Scientific Productivity," *Social Studies of Science*, vol. 10, no. 2, pp. 163–179, May 1980. [Online]. Available: http://dx.doi.org/10.1177/030631278001000203

[6] P. D. Allison, J. S. Long, and T. K. Kraze, "Cumulative advantage and inequality in science," *Ame. Sociological Review*, vol. 47, no. 5, pp. 615–625, 1982.

[7] A. Klamer and H. P. Van Dalen, "Attention and the art of scientific publishing," *Journal of Economic Methodology*, vol. 9, no. 3, pp. 289–315, 2002. [Online]. Available: http://www.klamer.nl/docs/2002_v9n3_JEM1.pdf

[8] E. S. Lang, P. C. Wyer, and R. B. Haynes, "Knowledge translation: closing the evidence-to-practice gap." *Annals of emergency medicine*, vol. 49, no. 3, pp. 355–363, 2007.

[9] Z. S. S. Morris, S. Wooding, and J. Grant, "The answer is 17 years, what is the question: understanding time lags in translational research." *Journal of the Royal Society of Medicine*, vol. 104, no. 12, pp. 510–520, 2011.

[10] J. Bednorz and K. Müller, "Possible high T_c superconductivity in the Ba–La–Cu–O system," *Z. Phys. B*, vol. 64, no. 2, pp. 189–193, Jun. 1986. [Online]. Available: http://dx.doi.org/10.1007/BF01303701

[11] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, "Theory of superconductivity," *Phys. Rev.*, vol. 108, pp. 1175–1204, Dec 1957. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRev.108.1175

[12] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences*, vol. 102, no. 46, pp. 16 569–16 572, Nov. 2005. [Online]. Available: http://dx.doi.org/10.1073/pnas.0507655102

[13] A. Mazloumian, Y.-H. Eom, D. Helbing, S. Lozano, and S. Fortunato, "How Citation Boosts Promote Scientific Paradigm Shifts and Nobel Prizes," *PLoS ONE*, vol. 6, no. 5, pp. e18 975+, May 2011.

[14] P. Bonacich, "Power and centrality: a family of measures," *The American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[15] R. Ghosh and K. Lerman, "A framework for quantitative analysis of cascades on networks," in *Proceedings of Web Search and Data Mining Conference (WSDM)*, February 2011.

[16] A. B. Kahn, "Topological sorting of large networks," *Communications of the ACM*, vol. 5, no. 11, pp. 558–562, 1962.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

[18] S. Redner, "Citation Statistics from 110 Years of Physical Review," *Physics Today*, vol. 58, no. 6, pp. 49–54, 2005. [Online]. Available: http://dx.doi.org/10.1063/1.1996475

[19] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in *KDD'08*, 2008, pp. 990–998.

[20] J. Tang, L. Yao, D. Zhang, and J. Zhang, "A combination approach to web user profiling," *ACM TKDD*, vol. 5, no. 1, pp. 1–44, 2010.

[21] J. Tang, J. Zhang, R. Jin, Z. Yang, K. Cai, L. Zhang, and Z. Su, "Topic level expertise search over heterogeneous networks," *Machine Learning Journal*, vol. 82, no. 2, pp. 211–237, 2011.

[22] J. Tang, D. Zhang, and L. Yao, "Social network extraction of academic researchers," in *ICDM'07*, 2007, pp. 292–301.

[23] P. Chen and S. Redner, "Community structure of the physical review citation network," Nov 2009, comments: 14 pages, 7 figures, 8 tables.

[24] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, "Microscopic theory of superconductivity," *Phys. Rev.*, vol. 106, pp. 162–164, Apr 1957. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRev.106.162

[25] M. K. Wu, J. R. Ashburn, C. J. Torng, P. H. Hor, R. L. Meng, L. Gao, Z. J. Huang, Y. Q. Wang, and C. W. Chu, "Superconductivity at 93 K in a new mixed-phase Y-Ba-Cu-O compound system at ambient pressure," *Physical Review Letters*, vol. 58, no. 9, 1987.

[26] F. C. Zhang and T. M. Rice, "Effective Hamiltonian for the superconducting Cu oxides," *Physical Review B*, vol. 37, no. 7, pp. 3759–3761, Mar. 1988. [Online]. Available: http://dx.doi.org/10.1103/PhysRevB.37.3759

[27] V. J. Emery, "Theory of high-t_c superconductivity in oxides," *Physical Review Letters*, vol. 58, no. 26, pp. 2794–2797, Jun. 1987. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.58.2794

[28] S. Iijima, "Helical microtubules of graphitic carbon," *Nature*, vol. 354, pp. 56–58, Nov. 1991. [Online]. Available: http://adsabs.harvard.edu/cgi-bin/nph-bib_query?bibcode=1991Natur.354...56I

[29] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, "Electric Field Effect in Atomically Thin Carbon Films," *Science*, vol. 306, no. 5696, pp. 666–669, 2004. [Online]. Available: http://www.sciencemag.org/cgi/content/abstract/306/5696/666

[30] K. S. Novoselov, D. Jiang, F. Schedin, T. J. Booth, V. V. Khotkevich, S. V. Morozov, and A. K. Geim, "Two-dimensional atomic crystals," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 30, pp. 10 451–10 453, Jul. 2005. [Online]. Available: http://dx.doi.org/10.1073/pnas.0502848102

[31] P. Chen, H. Xie, S. Maslov, and S. Redner, "Finding scientific gems with google's PageRank algorithm," *Journal of Informetrics*, vol. 1, no. 1, pp. 8–15, Jan. 2007. [Online]. Available: http://arxiv.org/abs/physics/0604130

[32] R. Ghosh, T.-T. Kuo, C.-N. Hsu, S.-D. Lin, and K. Lerman, "Time-aware ranking in dynamic citation networks," in *COMMPER 2011: Mining Communities and People Recommendations, Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, December 2011, pp. 373 –380.

[33] H. Sayyadi and L. Getoor, "Future rank: Ranking scientific articles by predicting their future PageRank," in *2009 SIAM International Conference on Data Mining (SDM09)*, 2009.

[34] C. Chen, "Searching for intellectual turning points: Progressive knowledge domain visualization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5303–5310, Apr. 2004. [Online]. Available: http://dx.doi.org/10.1073/pnas.0307513100

[35] S. Arbesman, *The Half-life of Facts: Why Everything We Know Has an Expiration Date*, first edition ed. Current Hardcover, Sep. 2012. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/159184472X

[36] K. Lerman and R. Ghosh, "Information contagion: an empirical study of spread of news on digg and twitter social networks," in *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, May 2010.

[37] S. Goel, D. J. Watts, and D. G. Goldstein, "The structure of online diffusion networks," in *Proceedings of the 13th ACM Conference on Electronic Commerce (EC 2012)*, 2012. [Online]. Available: http://5harad.com/papers/diffusion.pdf

[38] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading behavior in large blog graphs," in *Proceedings of 7th SIAM International Conference on Data Mining (SDM)*, Apr. 2007. [Online]. Available: http://arxiv.org/abs/0704.2803

[39] S. Myers and J. Leskovec, "Clash of the contagions: Cooperation and competition in information diffusion," in *Proceedings of ICDM*, 2012.

APPENDIX

Proof: **Theorem 1**

The proof is divided into two parts.

Part 1. Prove that let $S = \big(\mathbb{E}_{j\sim C'_t}[\phi_C(j)] - \mathbb{E}_{j\sim \widetilde{C}'_t}[\phi_C(j)]\big)$, we have

$$\big(\Phi_t(C') - \Phi_t(\widetilde{C'})\big) = \big(\mathbb{E}_{j\sim C'_t}[\phi_{C'}(j)] - \mathbb{E}_{j\sim \widetilde{C}'_t}[\phi_{C'}(j)]\big) > 0 \Longrightarrow S > 0.$$

by following the reasoning similar to that described in Section III-C.

Part 2. Prove

$$\Pr\left(\left|S - \big(\mathbb{E}_{j\sim C_t}[\phi_C(j)] - \mathbb{E}_{j\sim \widetilde{C}_t}[\phi_C(j)]\big)\right| > \varepsilon\right) < 2e^{-2\varepsilon^2|C'_t|}\frac{1}{4\varepsilon|C'_t|}\sqrt{\frac{\pi}{2|C'_t|}} + 2e^{-2\varepsilon^2|\widetilde{C}'_t|}\frac{1}{4\varepsilon|\widetilde{C}'_t|}\sqrt{\frac{\pi}{2|\widetilde{C}'_t|}}.$$

by applying the Hoeffding's inequality. ■