Inspiration or Preparation?
Explaining Creativity in Scientific Enterprise

Xinyang Zhang[†] Dashun Wang[‡] Ting Wang[†]
[†]Lehigh University, xizc15@lehigh.edu, inbox.ting@gmail.com
[‡]Northwestern University, dashunwang@gmail.com

December 6, 2016

It is the function of creative man to perceive and to connect the seemingly unconnected.

William Plomer

Abstract

Human creativity is the ultimate driving force behind scientific progress. While the building blocks of innovations are often embodied in existing knowledge, it is creativity that blends seemingly disparate ideas. Existing studies have made striding advances in quantifying creativity of scientific publications by investigating their citation relationships. Yet, little is known hitherto about the underlying mechanisms governing scientific creative processes, largely due to that a paper's references, at best, only partially reflect its authors' actual information consumption. This work represents an initial step towards fine-grained understanding of creative processes in scientific enterprise. In specific, using two web-scale longitudinal datasets (120.1 million papers and 53.5 billion web requests spanning 4 years), we directly contrast authors' information consumption behaviors against their knowledge products. We find that, of 59.0% papers across all scientific fields, 25.7% of their creativity can be readily explained by information consumed by their authors. Further, by leveraging these findings, we develop a predictive framework that accurately identifies the most critical knowledge to fostering target scientific innovations. We believe that our framework is of fundamental importance to the study of scientific creativity. It promotes strategies to stimulate and potentially automate creative processes, and provides insights towards more effective designs of information recommendation platforms.

1 Introduction

Among the many propulsions behind scientific progress, one stands out for its magnificent, yet intangible force - human creativity [17, 24]. While the building blocks of innovations are often embedded in existing knowledge, it is creativity that blends seemingly disparate concepts, ideas and theories [15, 11]. Indeed, even a theory as revolutionary as Einstein's special relativity essentially reconciles Newtonian mechanics and Maxwell's electromagnetic theory.

Since Plato's time [21], numerous studies in psychology, cognitive science and philosophy have offered a plethora of theories to explain human creativity [6, 13]. Despite their importance, today we still lack quantitative understanding of such phenomena. Thanks to prolific scientific publication archives (e.g., Dblp[1], PubMed[2], WoS[3]), we are now equipped with the lens to study creativity in *scientific enterprise* with unprecedented precision. Existing studies have focused on quantitatively gauging scientific papers' creativity by investigating their citation relationships [26, 8, 16]. These studies offer convincing evidences that a paper's

[1] http://dblp.uni-trier.de
[2] http://www.ncbi.nlm.nih.gov/pubmed/
[3] http://wokinfo.com

creativity is measurable by examining how it blends originally disconnected knowledge. Yet, little is known hitherto about the underlying mechanism governing this creative process.

This work represents an initial step towards fine-grained understanding of creativity in scientific enterprise. We argue that, to understand the mechanism underlying scientific creative processes, solely relying on papers' citation relationships is insufficient. After all, a paper's references, at best, only partially reflect its authors' actual information consumption behaviors. First, the references may not include the most critical literature. Second, the references may not provide a holistic view of all the literature inspiring the authors. Finally, to understand the correlation between information consumption and knowledge production, it is imperative to characterize their temporal dynamics; however, the references alone do not indicate when the cited literature were actually consumed by the authors. All these limitations highlight a fundamental gap between reality and perception in current studies.

We overcome these limitations by directly contrasting authors' information consumption behaviors with their knowledge products. In specific, using two web-scale, longitudinal datasets, Microsoft Academic Graph (120.1 million papers across all scientific fields) and Indiana University Click (53.5 billion web requests spanning over 4 years), we conduct a systematic study on creative processes in scientific enterprise. Even though varied privacy and technology constraints preclude the possibility of tracking information consumption and knowledge production at an individual level, by studying their correlation at an organization level, we find remarkable predictability in scientific creative processes: of 59.0% papers across all scientific fields, 25.7% of their creativity can be readily explained by information consumed by their potential authors.

Moreover, leveraging these findings, we develop a predictive framework that captures the impact of authors' information consumption over their future knowledge products. Using the aforementioned datasets as an exemplary case, we demonstrate that our framework is able to accurately identify the most critical knowledge to fostering target scientific innovations.

To our best knowledge, this work is among the first to study scientific creative processes within the context of information consumption. We believe that the proposed creativity metrics, in synergy with existing measures (e.g., citation counts), lead to more comprehensive understanding of scientific publications' merits. We also believe that our framework is of fundamental importance to the study of human creativity in general. Foreseeably, it promotes strategies to stimulate and potentially automate creative processes, and provides significant insights towards more effective designs of information recommendation platforms.

The remainder of the paper proceeds as follows. § 2 surveys relevant literature; § 3 describes the datasets used in our study; § 4 presents a general creativity definition that subsumes existing ones; § 5 explores the predictability in scientific creative processes within the context of information consumption; § 6 details our prediction framework and develops efficient optimization algorithms; § 4, § 5 and § 6 all conclude with empirical studies of the proposed models and algorithms; the paper is summarized in § 7.

2 Related Work

In this section, we review four categories of related work: assessment of creativity, scientific impact prediction, map of science, and computational creativity.

Despite numerous qualitative studies on creativity pertaining to various disciplines: psychology, cognitive science, economics and philosophy [17, 6, 4, 29, 12, 13], it is only after the proliferation of scientific publication archives that it is feasible to quantitatively study creativity in scientific enterprise. One active line of inquiry is to develop meaningful creativity metrics. Uzzi et al. [26] proposed to measure a paper's creativity as atypical pairwise combinations of its referenced work; Fleming [12] proposed to gauge a patent's novelty using new combinations of patents in its references. This line of work offers convincing evidence that scientific work's creativity is measurable by investigating how it blends originally disconnected knowledge. However, little is known hitherto about the mechanism that triggers such connections. To our best knowledge, this work is among the first to directly bridge this gap by studying creativity within the context of information consumption.

Meanwhile, another line of work has focused on predicting a paper's long-term impact (primarily measured by its citation count) in its early stages [28, 3, 9, 18] using various semantic features (e.g., author, content,

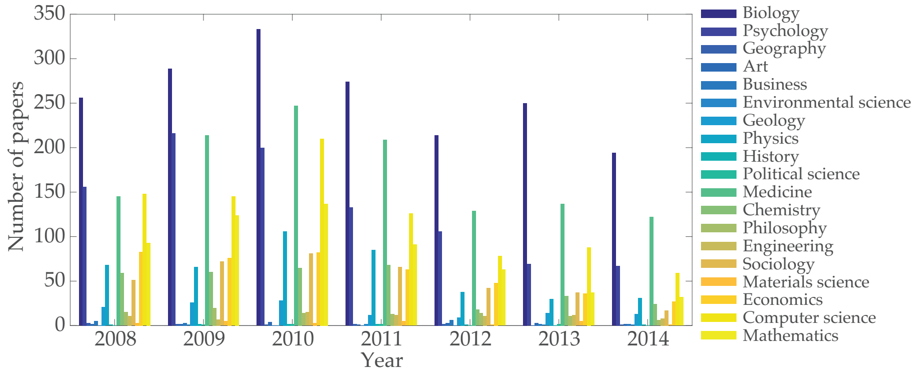

Figure 1: Number of Indiana University publications in each scientific field per year.

venue). These studies are complementary to this work in that the useful semantic features can be integrated into our framework to train microscopic (e.g., author-, content-, and venue-specific) creativity models.

Another use of papers' reference relationships is to create citation-based *maps of science* or *knowledge flow maps* [22, 14], which help categorize science and understand papers' trans-disciplinary impact. However, these insights do not help explain creativity of individual scientific work.

Finally, this work is related to the broad area of computational creativity [30, 7], which focuses on developing artificial intelligence models that exhibit or generate creativity (e.g., problem solving [23], visual creativity [5], and linguistic creativity [27]). In contrast, this work focuses on understanding and modeling creative processes in scientific enterprise. However, incorporating these intelligence models to enhance the predictive power of our framework would be one promising direction.

3 Data

Next we describe the datasets used in our empirical study.

3.1 Raw Data

We used two web-scale, longitudinal datasets, corresponding to information consumption and knowledge production of scientific creative processes, respectively.

Information Consumption

At present, the most comprehensive dataset that captures information consumption in scientific enterprise is perhaps the web traffic generated by researchers, reflecting how they request and access online resources (e.g., online publication archives). Therefore, in our study, we used the Indiana University Click Dataset [19] (Click), which constitutes about 53.5 billion web requests initiated by researchers at Indiana University from 09/2006 to 05/2010. This anonymized dataset was collected by applying a Berkeley Packet Filter to the web traffic passing through the border router of Indiana University and matching all the traffic containing a HTTP GET request.

Each request consists of the following fields: ⟨*timestamp, requested url, referring url, agent, flag*⟩, where "*agent*" indicates whether the user agent was a browser or a bot, while "*flag*" indicates whether the request was generated inside or outside of Indiana University. All incoming or bot-generated requests have been filtered.

Knowledge Production

Meanwhile, to capture knowledge production in scientific enterprise, we used the Microsoft Academic Graph Dataset [25] (Mag). As of 11/06/2015, this dataset consists of 120.9 million papers published in 24,843

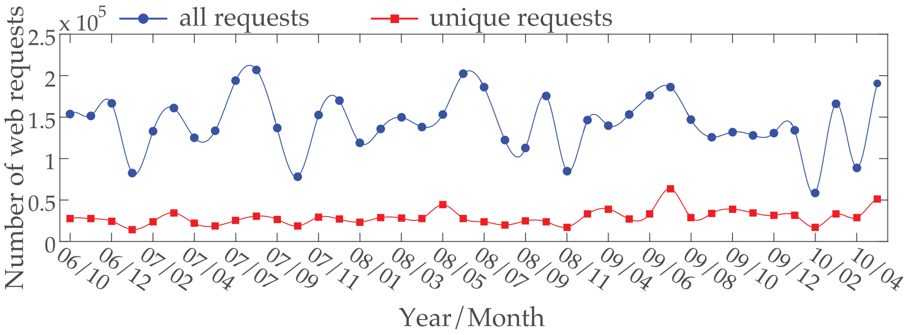

Figure 2: Total number of requests and number of unique requests per month.

venues across all scientific fields.

In a nutshell, the Mag dataset is a web-scale entity graph comprising scientific publication records, citation relationships between publications, as well as their authors, author affiliations, publication venues, keywords and fields-of-study (i.e., *topics*). In particular, all the topics form a four-level hierarchy, with the highest and lowest levels corresponding to disciplines (e.g., "*Computer science*") and specific subjects (e.g., "*Decision tree*"), respectively. Each keyword is associated with one topic in the hierarchy.

3.2 Preprocessing

Next, to match corresponding information consumption and knowledge production data, we correlate the Click and Mag datasets as follows.

In the Mag dataset, we identified all the papers that have at least one author affiliated with Indiana University and were published during the period from 2007 to present, resulting in a collection of 24,399 papers. Figure 1 illustrates the number of papers in each scientific field from 2008 to 2014. It is noted that both the number and composition of papers vary significantly on a yearly basis.

Further, in the Click dataset, we identified all the web requests that ask for papers in the Mag dataset by matching URLs embedded in the requests with URLs of the papers in the Mag dataset. The resulting dataset consists of 5.8 million requests for 4.6 million papers (i.e., unique requests). Figure 2 illustrates the total number of requests and the number of unique ones per month from 09/2006 to 05/2010. Note that while the total number of requests fluctuates wildly.

4 Measurement & Quantification

In this section, we present a general creativity definition and apply it to the aforementioned datasets to empirically study scientific publications' creativity. We start by introducing a set of fundamental concepts and assumptions used throughout the paper.

4.1 Preliminaries

We refer to a scientific publication as a "paper". We assume that each paper k is described by a tuple $\langle t_k^p, \mathcal{K}_k, \mathcal{C}_k \rangle$, representing its publishing time, keywords and references, respectively. Note that this model can be easily extended by including additional information (e.g., abstract, full text and publishing venue). Further, we refer to each web request with respect to any paper in online publication archives as a "reading". We assume that a reading is described by a tuple $\langle k, t_k^r \rangle$, denoting the reading paper and the time of reading respectively. Note that with respect to a given paper k which is requested ("read") multiple times, we consider the median of its reading time as t_k^r.

Let $\mathcal{P}_{t,t'}$ denote the papers published within the time period from t to t'. In particular, $\mathcal{P}_{,t}$ represents all the papers published till t. Similarly, denote by $\mathcal{Q}_{t,t'}$ the set of papers read during the time window

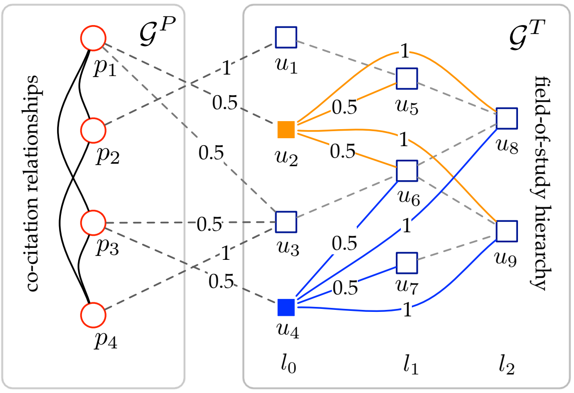

Figure 3: A schematic example of paper and field-of-study heterogeneous network.

from t to t'. Given that (i) for a large number of papers in the MAG dataset, only their publishing years are specified and (ii) even with finer grained timestamps, a paper's publication often deviates from when it is actually finished, we thus use "year" as the default time granularity in our study. Thus, with a little abuse of notations, we use t to denote both a timestamp and a one-year-long time window. For example, $\mathcal{P}_t$ represents the set of papers published in year t.

4.2 A General Creativity Definition

A variety of creativity metrics have been proposed in literature (e.g., [12, 26]), all premised on the same intuition: a paper's creativity should be measured by how it blends originally disconnected knowledge. To capture this intuition, one needs to consider two factors: (i) the "disconnect" (denoted by $d_{i,j}$) of knowledge represented by two papers (i, j), and (ii) the "rarity" (denoted by $r_{i,j}$) that the knowledge of (i, j) has been connected in previously published papers. We define the product of disconnect and rarity,

$$\varphi_{i,j} = d_{i,j} \cdot r_{i,j} \tag{1}$$

as the creativity score of co-citation of (i, j). We then introduce the following creativity metric.

Definition 1 (Creativity)**.** *A paper k's overall creativity ϕ_k is the aggregation of creativity scores contributed by all its reference pairs:*

$$\phi_k = \ell(\{\varphi_{i,j}\}_{i,j \in \mathcal{C}_k}) \tag{2}$$

where $\ell(\cdot)$ represents the aggregation function (e.g., average, percentile, maximum).

We require that $\ell(\cdot)$ is non-decreasing: if $\forall i, j \in \mathcal{C}_k$, $\varphi_{i,j} \le \varphi'_{i,j}$, then $\ell(\{\varphi_{i,j}\}_{i,j\in\mathcal{C}_k}) \le \ell(\{\varphi'_{i,j}\}_{i,j\in\mathcal{C}_k})$. We remark that this definition subsumes a number of creativity metrics in literature. For example, the metrics in [12, 26] only consider the rarity factor, while the measures in [8] only reflect the disconnect factor.

4.3 Rarity versus Disconnect

Next we discuss the concrete instantiation of $\varphi_{i,j}$. A variety of forms are possible. For example, $r_{i,j}$ can be gauged using the frequency of co-occurrences of (i, j)'s publishing venues [26]. Similar definitions may be given based on papers' author or affiliation information. In our study, we define (i, j)'s creativity score within the context of their associated topics.

Recall that each keyword in the MAG dataset is associated with one topic. We define paper k's topics (denoted by $\mathcal{T}_k$) as the aggregation of topics associated with its keyword set $\mathcal{K}_k$. For example, the paper "*Fast algorithms for mining association rules*" [1] is associated with the keyword of "*association rule*", which is mapped to the topic of "*Association rule learning*".

Let $\mathcal{P}$, $\mathcal{K}$ and $\mathcal{T}$ be the sets of papers, keywords and topics, respectively. We introduce a heterogeneous network to describe papers' semantic relationships within the context of their topics, as shown in Figure 3. The paper network $\mathcal{G}^P = (\mathcal{P}, \mathcal{E}^P)$ captures papers' co-citation relationships; each edge $(i, j) \in \mathcal{E}^p$ indicates that two papers $i, j \in \mathcal{P}$ are both referenced by some other paper(s). We now introduce the concept of rarity.

Definition 2 (Rarity)**.** *The rarity of co-citation of (i, j) till year t, $r^t_{i,j}$, is defined as follows:*

$$r^t_{i,j} = \frac{1}{1 + \log_2(c^t_{i,j} + 1)} \tag{3}$$

where $c^t_{i,j}$ is the number of co-citations of (i, j) till year t.

Meanwhile, the topic network $\mathcal{G}^T = (\mathcal{T}, \mathcal{E}^T)$ encodes the four-level field-of-study hierarchy; a topic u may have one or more parent topics $\mathcal{H}^l_u$ at each higher level l. For example, in Figure 3, u_2 has two parents u_5, u_6 at level l_1 and two parents u_8, u_9 at level l_2. Given $u' \in \mathcal{H}^l_u$, the weight $w_{u,u'}$ of edge $(u, u') \in \mathcal{E}^T$ indicates the confidence that u is a sub-topic of u', with $\sum_{u' \in \mathcal{H}^l_u} w_{u,u'} = 1$.

We define the disconnect of (i, j) based on their topics $(\mathcal{T}_i, \mathcal{T}_j)$. Let us start with the simplest case that $|\mathcal{T}_i| = |\mathcal{T}_j| = 1$, i.e., $\mathcal{T}_i = \{u\}$ and $\mathcal{T}_j = \{v\}$. Without loss of generality, we assume that (u, v) lie at the same level of topic hierarchy. The similarity of (u, v) is the aggregation of level-wise similarity of (u, v) and their parents. Three properties are desirable: (i) if (u, v) are identical, their similarity should be 1; (ii) the similarity $s^l_{u,v}$ of $(\mathcal{H}^l_u, \mathcal{H}^l_v)$ is discounted with respect to l; and (iii) $s^l_{u,v}$ is counted only if $\{(\mathcal{H}^{l'}_u, \mathcal{H}^{l'}_v)\}_{l'<l}$ are not identical.

Thus, for $l = 0$ to 3 ($\mathcal{H}^0_u = \{u\}$, $\mathcal{H}^0_v = \{v\}$), we compute the level-l similarity as: $s^l_{u,v} = \sum_{x \in \mathcal{H}^l_u \cap \mathcal{H}^l_v} \min(w_{u,x}, w_{v,x})$. Then the overall similarity of u, v is given as:

$$s_{u,v} = \sum_{l=0}^{3} \max\left(1 - \sum_{l'=0}^{l-1} s^{l'}_{u,v}, 0\right) \cdot s^l_{u,v} \cdot \sigma^l$$

where the first term represents the "budget" after reaching level l and σ is the "discount". It can be verified that this definition fulfills all three desirable properties.

For example, in Figure 3, we compute the level-wise similarity of (u_2, u_4): $s^0_{u_2,u_4} = 0$, $s^1_{u_2,u_4} = \min(w_{u_2,u_6}, w_{u_4,u_6}) = 0.5$, $s^2_{u_2,u_4} = \sum_{u=u_8,u_9} \min(w_{u_2,u}, w_{u_4,u}) = 1$. The overall similarity of (u_2, u_4) is given as: $s_{u_2,u_4} = 0 + 0.5 * 0.8 + (1 - 0.5) * 1 * 0.8^2 = 0.72$, if $\sigma = 0.8$.

Now we are ready to introduce the disconnect metric:

Definition 3 (Disconnect)**.** *The disconnect of two reference papers (i, j) is defined as the average dissimilarity of their topics $(\mathcal{T}_i, \mathcal{T}_j)$:*

$$d_{i,j} = 1 - \frac{1}{|\mathcal{T}_i||\mathcal{T}_j|} \sum_{u \in \mathcal{T}_i, v \in \mathcal{T}_j} s_{u,v} \tag{4}$$

4.4 Empirical Study

Next we apply the metrics above in our empirical study. Due to space limitations, here we only show results obtained using papers published in 2011. Similar phenomena are observed regarding papers in other years.

Rarity, Disconnect and Creativity

With a little abuse of notations, let $\mathcal{P}$ denote all the papers published in 2011 by Indiana University. We first measure the rarity and disconnect of all the reference pairs of $\mathcal{P}$, i.e., $\{(i, j) \in \mathcal{C}_k\}_{k \in \mathcal{P}}$. The results are shown in Figure 4.

Figure 4(a) shows the cumulative distribution of (i, j)'s rarity $r_{i,j}$. Note that around 60% pairs have zero co-citation before 2011. In Figure 4(b), we further measure the conditional distribution of (i, j)'s disconnect $d_{i,j}$ on its rarity $r_{i,j}$. The conditional distribution $Pr(d_{i,j}|r_{i,j})$ is fairly similar irrespective to varying

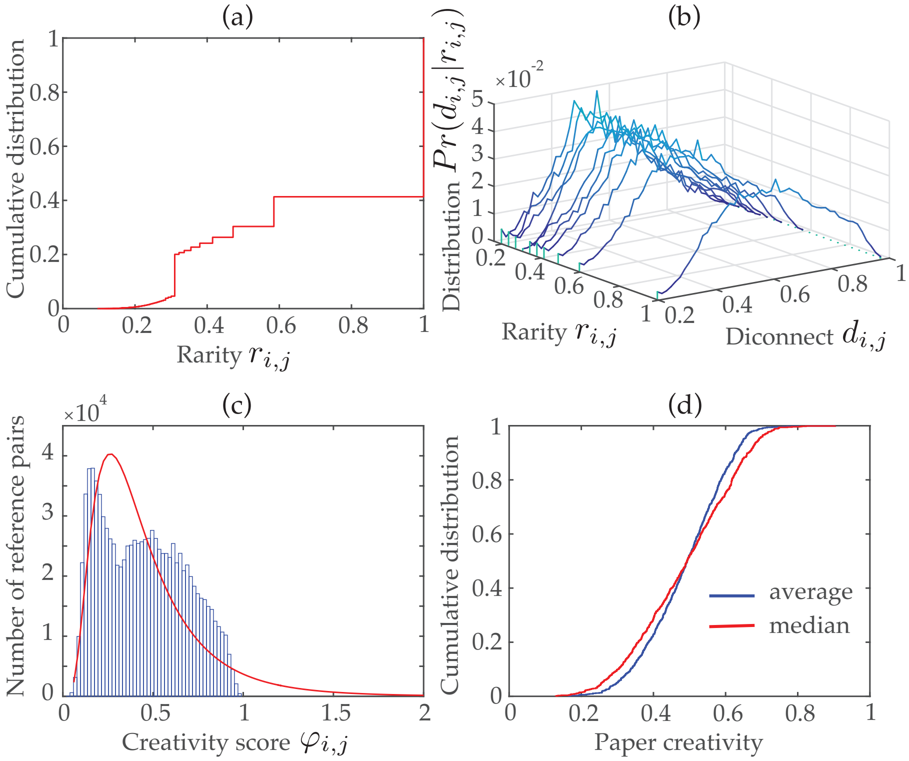

Figure 4: (a) Cumulative distribution of reference pairs' rarity; (b) Distribution of reference pairs' disconnect conditional on their rarity; (c) Histogram of reference pairs' creativity scores; (d) Cumulative distribution of papers' creativity (with average and median as aggregation functions).

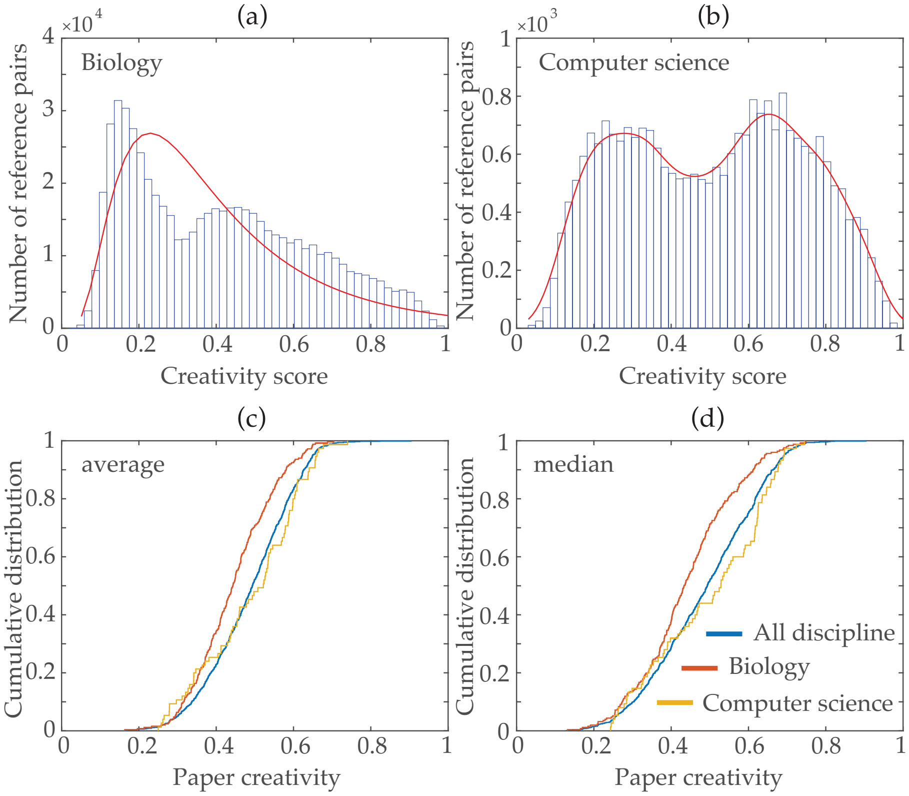

Figure 5: (a)-(b) Histogram of reference pairs' creativity scores (*Biology* and *Computer science*); (c)-(d) Cumulative distribution of papers' creativity (average and median).

$r_{i,j}$, implying that disconnect and rarity are two critical, but complementary factors of creativity. Then, according to Eqn.(1), we integrate (i, j)'s rarity and disconnect to compute its creativity score $\varphi_{i,j}$. As shown in Figure 4(c), the histogram of $\varphi_{i,j}$ roughly follows a lognormal distribution. It can be intuitively explained by that most reference pairs represent modestly "common" combinations, while both extremely "cliché" and "atypical" combinations are rare. Finally, we measure the creativity of papers in $\mathcal{P}$ by aggregating their reference pairs' scores. We consider both average and median as the aggregation function $\ell(\cdot)$, with results depicted in Figure 4(d). Clearly, most of the papers show moderate creativity, which is consistent with the results reported in prior studies [26].

Discipline-Specific Patterns

We further investigate the discipline-specific patterns of creativity measures. Figure 5(a)-(b) demonstrate the histogram of (i, j)'s creativity score $\varphi_{i,j}$ in the disciplines of *Biology* and *Computer science* respectively. Interestingly, one can observe that $\varphi_{i,j}$ roughly follows a lognormal distribution in Figure 5(a). Meanwhile, in Figure 5(b), $\varphi_{i,j}$ apparently follows a bimodal distribution, peaking at both low and high creativity scores. Such phenomena may be explained by that compared with *Biology*, *Computer science* is a relatively "engineering" discipline, featuring more frequent fusion of originally disconnected knowledge. As shown in Figure 5(c)-(d), this difference indeed leads to the disparate creativity distributions of papers published in these two disciplines: in general, the papers in *Computer science* demonstrate higher creativity.

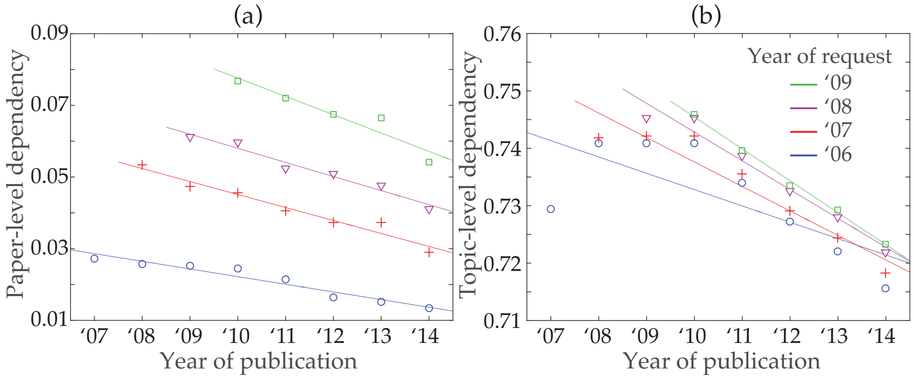

Figure 6: Paper- and topic-level dependency of knowledge production on information consumption.

5 Anatomy of Creativity

Equipped with the metrics introduced above, we are able to quantitatively assess the creativity of a paper, an author, an institution or even a discipline. Next we will explore the *predictability* of creativity by directly contrasting authors' raw information consumption behaviors against their knowledge products. We intend to answer the following key question:

Which part of a paper's creativity can be explained by papers read by its authors?

5.1 Production-Consumption Dependency

We begin with validating the dependency of authors' publishing papers on their reading papers. In specific, for the papers published in year t, $\mathcal{P}_t$, we compare the set of prior work $\mathcal{C}_t$ referenced by $\mathcal{P}_t$ (i.e., $\mathcal{C}_t = \cup_{k \in \mathcal{P}_t} \mathcal{C}_k$) against the set of reading papers $\mathcal{Q}_{t'}$ in an earlier year t' ($t' < t$). We measure the dependency of $\mathcal{P}_t$ on $\mathcal{Q}_{t'}$ at two distinct levels.

Paper-Level Dependency

We first directly compare reference set $\mathcal{C}_t$ against reading set $\mathcal{Q}_{t'}$. In specific, we measure the paper-level dependency of $\mathcal{P}_t$ on $\mathcal{Q}_{t'}$ by computing the Jaccard's coefficient of $\mathcal{C}_t$ and $\mathcal{Q}_{t'}$:

$$\frac{|\mathcal{C}_t \cap \mathcal{Q}_{t'}|}{\min\{|\mathcal{C}_t|, |\mathcal{Q}_{t'}|\}}$$

Figure 6(a) illustrates the paper-level dependency in our datasets. As t varies from 2007 to 2014, we measure the dependency of $\mathcal{P}_t$ on $\mathcal{Q}_{t'}$ for $t'(t' < t)$ ranging from 2006 to 2009.

It is observed that this paper-level dependency demonstrates interesting temporal dynamics. First, given reading set $\mathcal{Q}_{t'}$ (i.e., fixed t'), the dependency of $\mathcal{P}_t$ on $\mathcal{Q}_{t'}$ decreases as t grows, implying that the impact of $\mathcal{Q}_{t'}$ gradually decays over time. Second, for given publication set $\mathcal{P}_t$ (i.e., fixed t), its dependency on $\mathcal{Q}_{t'}$ increases with t', implying that more recent reading papers exert more influence over future publications.

Topic-Level Dependency

We then examine the dependency of publication set $\mathcal{P}_t$ on reading set $\mathcal{Q}_{t'}$ at the topic level.

In specific, following the definition of topic disconnect in Eqn.(4), we compute the topic-level dependency of $\mathcal{P}_t$ on $\mathcal{Q}_{t'}$ as the average "connectedness" of papers in $\mathcal{P}_t$ and $\mathcal{Q}_{t'}$:

$$1 - \frac{\sum_{i \in \mathcal{P}_t} \sum_{j \in \mathcal{Q}_{t'}} d_{i,j}}{|\mathcal{P}_t||\mathcal{Q}_{t'}|}$$

The measurement results are illustrated in Figure 6(b). We have the following observations.

The topic-level dependency shows pattens similar to the paper-level dependency: (i) the impact of reading papers over future publications decays over time; (ii) more recent reading papers exert stronger influence. Nevertheless, compared with the paper-level dependency, the topic-level dependency seems less "stratified" in that adjacent years show more resemble patterns. For instance, the dependency with respect to the reading papers in 2006 and 2007 shows strong similarity. Such phenomena can be explained by that authors' interests in particular topics are more stable than their interests in concrete papers.

5.2 A Dichotomic Theory of Creativity

The study on paper- and topic-level dependency above offers empirical evidences for our premise that authors' future publications are heavily influenced by prior literature which they have read. This also hints that it is conceivable to answer the *creativity prediction* question posed at the beginning of this section.

Impact of Information Consumption

To answer this question, we introduce a mathematical model to quantify the impact of information consumption in creative processes. To motivate the rationale behind our model, let us consider the following concrete example of three papers by three disjoint groups of authors.

- i: "*Fast algorithms for mining association rules*". R. Agrawal and R. Srikant. *VLDB '94*.
- x: "*Frequent subgraph discovery*". M. Kuramochi and G. Karypis. *ICDM '01*.
- j: "*SPIN: mining maximal frequent subgraphs from graph databases*". J. Huan, W. Wang, J. Prins and J. Yang. *KDD '04*.

Here paper j differs from i significantly. However, after reading paper x, one is probably able to draw the connection from i to j as "*frequent pattern*" → "*frequent subgraph pattern*" → "*maximal frequent subgraph pattern*". Informally, reading x offers the opportunity to bridge the knowledge gap between i and j.

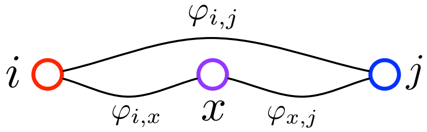

Figure 7: Impact of x on the creativity required to connect (i, j): $\Delta_x^{i,j} = \varphi_{i,j} - \min(\varphi_{i,j}, \max(\varphi_{i,x}, \varphi_{x,j}))$.

We capture this intuition using the metric of creativity score defined in Eqn.(1). The term $\varphi_{i,j}$ can be interpreted as the "difficulty" of connecting (i, j), while the exposure to x may potentially reduce this difficulty. That is, instead of directly connecting (i, j), one may now first connect i to x, and then link x to j, as illustrated in Figure 7.

Formally, we introduce a metric to describe the impact of paper x on the creativity required to connect (i, j):

$$\Delta_x^{i,j} = \varphi_{i,j} - \min(\varphi_{i,j}, \max(\varphi_{i,x}, \varphi_{x,j})) \tag{5}$$

where the term $\max(\varphi_{i,k}, \varphi_{k,j})$ represents the difficulty of connecting (i, x) and (x, j).

Note that this definition can be generalized to the case that n papers $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ collectively bridge the gap of (i, j) by creating an n-hop "path" between them. Let $\varphi_{\boldsymbol{x}} = \max_{l=1}^{n-1} \varphi_{x_l, x_{l+1}}$. We may define the impact of $\boldsymbol{x}$ to (i, j) as: $\Delta_{\boldsymbol{x}}^{i,j} = \varphi_{i,j} - \min(\varphi_{i,j}, \max(\varphi_{i,x_1}, \varphi_{\boldsymbol{x}}, \varphi_{x_n,j}))$. Due to space limitations, in the following discussion we focus on the case of $n = 1$.

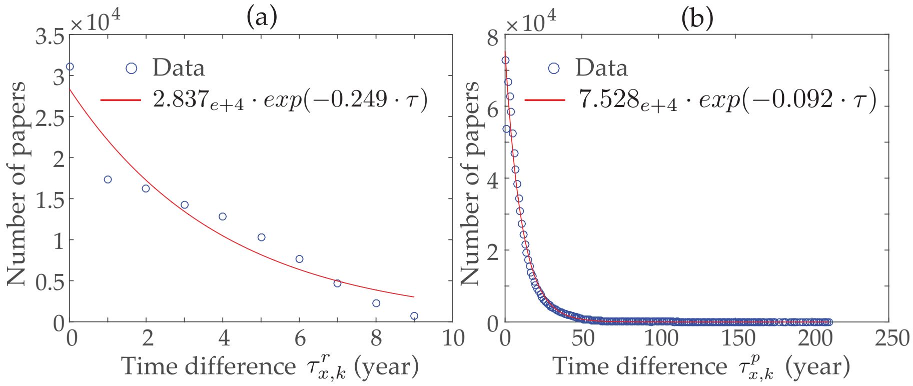

Figure 8: Temporal dynamics of information consumption-production. (a) Histogram of consumption-publication temporal interval $\tau^r_{x,k}$; (b) Histogram of publication-publication temporal interval $\tau^p_{x,k}$.

Preparation versus Inspiration

We are now ready to perform an anatomy of the creativity $\varphi_{i,j}$ required to connect the knowledge represented by two papers (i, j). For simplicity, consider the case that during this creative process, the authors read a single paper x. We divide $\varphi_{i,j}$ into two complementary parts.

- "Preparation" - the marginal reduction in the difficulty of connecting (i, j), due to the authors' reading, which is quantified by $\Delta^{i,j}_x$.
- "Inspiration" - the part of creativity that cannot be explained by the authors' reading, which is quantified by $(\varphi_{i,j} - \Delta^{i,j}_x)$.

Since a given paper blends the knowledge in all its references, a dichomatic theory naturally follows. That is, a paper's creativity reflects a superimpose of both effects:

$$\boxed{\textbf{Creativitiy} = \textbf{Preparation} + \textbf{Inspiration}}$$

Formally, consider a paper k with $\mathcal{C}_k$ as its reference set. Assume that k's authors have read a set of papers $\mathcal{Q}$ during their creative process. As given in Eqn.(2), k's overall creativity is assessed by: $\phi_k = \ell(\{\varphi_{i,j}\}_{i,j\in\mathcal{C}_k})$. For each pair $i, j \in \mathcal{C}_k$, we identify a reading paper in $\mathcal{Q}$ that maximally impacts $\varphi_{i,j}$ and compute its impact as:

$$\Delta^{i,j}_{\mathcal{Q}} = \max_{x\in\mathcal{Q}} \Delta^{i,j}_x$$

Definition 4 (Enabler)**.** *For reference pair $i, j \in \mathcal{C}_k$, the paper x^* in reading set $\mathcal{Q}$ that maximally impacts $\varphi_{i,j}$, $x^* = \arg\max_{x\in\mathcal{Q}} \Delta^{i,j}_x$ is called an enabler.*

Thus, after discounting the preparation quantity embodied in $\mathcal{Q}$, the inspiration of k, χ_k, is measured as:

$$\chi_k = \ell(\{\varphi_{i,j} - \Delta^{i,j}_{\mathcal{Q}}\}_{i,j\in\mathcal{C}_k})$$

while the preparation quantity ψ_k is computed as the difference of creativity and inspiration measures:

$$\psi_k = \ell(\{\varphi_{i,j}\}_{i,j\in\mathcal{C}_k}) - \ell(\{\varphi_{i,j} - \Delta^{i,j}_{\mathcal{Q}}\}_{i,j\in\mathcal{C}_k})$$

Temporal Dynamics

In the creativity theory above, the definition of $\mathcal{Q}$ is critical for accurately quantifying preparation and inspiration. While one may regard all the papers read by given authors in the past as $\mathcal{Q}$, this simplification ignores the temporal dynamics of information production-consumption dependency observed in § 5.1, resulting in an overestimation of preparation quantities. Here we will address this issue.

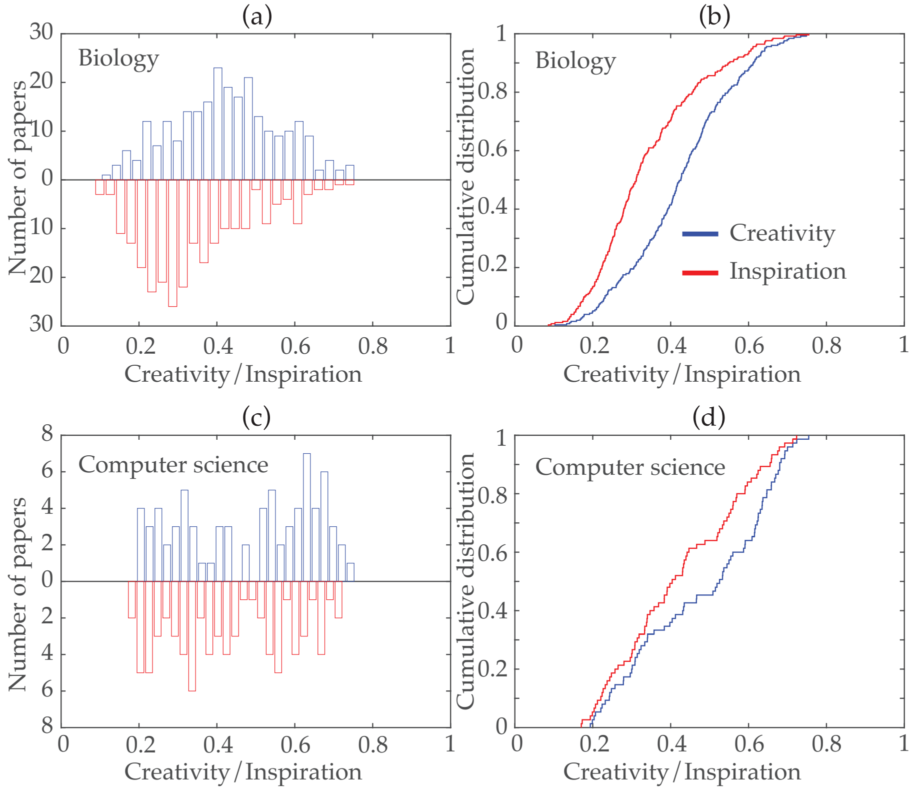

Figure 9: (a)-(c) Papers' creativity and inspiration in the disciplines of *Biology* and *Computer science*; (b)-(d) Cumulative distribution of papers' creativity and inspiration in the disciplines of *Biology* and *Computer science*.

It is extremely difficult to directly assess the temporal dynamics of a reading paper x's impact over a future publishing paper k. However, by studying in general how authors access papers in the intersection of reference and reading sets (i.e., $\mathcal{C}_k \cap \mathcal{Q}$), we are able to build a surrogate temporal dynamics model.

Specifically, for $x \in \mathcal{C}_k \cap \mathcal{Q}$, we examine x's reading time (t^r_x) and k's publishing time (t^p_k). The difference ($\tau^r_{x,k} = t^p_k - t^r_x$) reflects the temporal interval between information consumption and knowledge production. Figure 8(a) shows that this interval follows a log-log distribution, which is consistent with the studies on papers' "aging" phenomena [10]. Let $Pr(\tau)$ represent this distribution. By discretizing and normalizing $Pr(\tau)$, we compute the probability that a reading paper influences a future paper published Δt years later, i.e., $m(\Delta t) = \int_{\tau=\Delta t}^{+\infty} Pr(\tau)\mathrm{d}\tau$.

Putting everything together, Algorithm 1 sketches how to compute a given paper k's preparation and inspiration quantities. For each reference pair $i, j \in \mathcal{C}_k$, the papers in reading set $\mathcal{Q}$ are ranked according to their impact to the creativity score of (i, j); the current top one x^* is picked with probability $m(\tau^r_{x^*,k})$; if x^* is not picked, it moves to the next most impactful one and repeats this process; finally, k's preparation and inspiration measures are computed by aggregating the creativity scores of all reference pairs.

5.3 Empirical Study

Next we apply this theory to explain the creativity of papers in our datasets.

Inspiration-Preparation Decomposition

Applying Algorithm 1, we first assess the preparation and inspiration quantities of papers published in 2011 (similar phenomena are observed in other years). In particular, we examine the papers in the disciplines of

Algorithm 1: Explaining creativity in scientific work

```
Input: publishing paper $k$, reading papers $\mathcal{Q}$
Output: preparation and inspiration of $k$
for each pair of $i, j \in \mathcal{C}_k$ do
    /* preparation and inspiration to connect $i, j$ */
    candidate references $\tilde{\mathcal{Q}} \leftarrow \mathcal{Q}$;
    while true do
        find $x^* = \arg\max_{x \in \tilde{\mathcal{Q}}} \Delta_x^{i,j}$;
        remove $x^*$ from $\tilde{\mathcal{Q}}$;
        /* temporal dynamics-based Bernoulli trial */
        pick $x^*$ with probability $m(\tau_{x^*,k}^r)$;
        if $x^*$ is picked then break
    compute $\Delta_{\mathcal{Q}}^{i,j} \leftarrow \Delta_{x^*}^{i,j}$;
/* aggregate at paper level */
$\chi_k \leftarrow \ell(\{\varphi_{i,j} - \Delta_{\mathcal{Q}}^{i,j}\}_{i,j \in \mathcal{C}_k})$;
$\psi_k \leftarrow \ell(\{\varphi_{i,j}\}_{i,j \in \mathcal{C}_k}) - \ell(\{\varphi_{i,j} - \Delta_{\mathcal{Q}}^{i,j}\}_{i,j \in \mathcal{C}_k})$;
```

Biology and *Computer science* separately.

Figure 9 contrasts a paper k's overall creativity ϕ_k and inspiration χ_k. Specifically, Figure 9(a)-(c) show the histograms of ϕ_k and χ_k regarding the papers in *Biology* and *Computer science*, respectively; Figure 9(b)-(d) further compare their cumulative distributions. It is observed in both cases, ψ_k accounts for a sizable portion of ϕ_k. Table 1 summarizes the impact of ψ_k. We examine the papers whose creativity measures drop after the preparation is taken into account. For example, of about 59.0% papers across all scientific fields, 25.7% of their creativity can be explained by information consumed by their authors. Also interestingly, the quantity of preparation varies with specific disciplines: it counts for 21.9% and 29.2% of creativity in the disciplines of *Computer science* and *Biology*, respectively. This variance may be explained by that in a more established discipline as *Biology*, authors need to ground their research in existing work more profoundly.

Discipline	**Papers w. ↓ (%)**	**Avg. ↓**	**Avg. ↓ (%)**
Biology	64.54%	0.132	29.18%
Computer science	48.00%	0.123	21.88%
All disciplines	58.96%	0.133	25.67%

Table 1. Impact of preparation over overall creativity.

Case Study

We further explore the practical use of this creativity theory. Here we report one concrete case.

The target paper k: "*Engaging online learners: the impact of web-based learning technology on college student engagement*" cites two papers i: "*A comprehensive look at online student support services for distance learners*" and j: "*Do computers enhance or detract from student learning?*". The reading paper x: "*The convergent and discriminant validity of NSSE scalelet scores*" reduces $\varphi_{i,j}$ from 0.661 to 0.495.

At a first glance, x is not relevant to either i or j. Yet, after manually examining the full text of these papers, one may notice the following connection: to build an online learning environment (i's knowledge), it is critical to effectively assess student engagement in this new environment (j's knowledge), while x indeed provides a metric to evaluate student engagement.

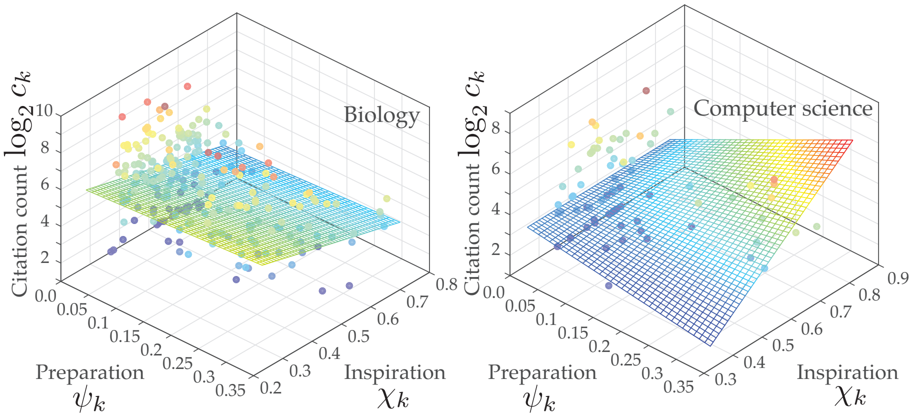

Figure 10: Relationships of preparation and inspiration with citation counts.

Relationships with Long-Term Impact

Finally, we investigate the relationships of preparation and inspiration with other important metrics of scientific work. In particular, we focus on a paper's citation count, the de facto metric of its long-term impact [28, 3, 9, 18].

Figure 10 plots the correlation of a paper k's preparation ψ_k and inspiration χ_k with its citation count c_k. Again, we examine papers in the disciplines of *Biology* and *Computer science* separately. In both cases, c_k is positivity correlated with ψ_k, implying that a paper grounded deeper into exiting work is better received by the research community. More interesting is the correlation of c_k and χ_k: it is positive in the case of *Computer science*, yet slightly negative in the case of *Biology*. This may be explained by that as a younger and more vibrant discipline, the *Computer science* community tends to be more welcoming to radical ideas that blend previously disconnected knowledge.

Clearly, the creativity, preparation and inspiration metrics provide a new perspective to assess scientific publications' merits. We envision that in synergy with other metrics (e.g., citation count), they may lead to more comprehensive understanding of long-term impact of scientific work.

6 Prediction of Enablers

In this section, levering the insights derived from our empirical study, we develop a predictive model that is able to identify the most promising enablers with respect to target papers. We then present efficient optimization algorithms for this model.

6.1 Problem Formulation

Without loss of generality, we start with the setting of a single target paper, and will discuss the extension to multiple target papers shortly. Let k denote the target paper, which connects a set of disparate knowledge, as represented by a set of references $\mathcal{C}_k$. Among all existing literature $\mathcal{S}$, we intend to find a minimum set of reading papers $\mathcal{A} \subseteq \mathcal{S}$ that maximally facilitate to connect $\mathcal{C}_k$.

More formally, our goal is to solve the following optimization problem:

$$\max_{\mathcal{A} \subseteq \mathcal{S}} \ell(\{\varphi_{i,j}\}_{i,j \in \mathcal{C}_k}) - \ell(\{\varphi_{i,j} - \Delta_{\mathcal{A}}^{i,j}\}_{i,j \in \mathcal{C}_k}) \tag{6}$$

with the constraint of $|\mathcal{A}| \leq \rho$ (ρ as the threshold).

Let $R(\mathcal{A}) \triangleq \ell(\{\varphi_{i,j}\}_{i,j \in \mathcal{C}_k}) - \ell(\{\varphi_{i,j} - \Delta_{\mathcal{A}}^{i,j}\}_{i,j \in \mathcal{C}_k})$ and $c(\mathcal{A}) \triangleq |\mathcal{A}|$. We can simplify this problem as:

$$\max_{\mathcal{A} \subseteq \mathcal{S}} R(\mathcal{A}) \quad \text{subject to} \quad c(\mathcal{A}) \leq \rho$$

We may interpret $R(\mathcal{A})$ and $c(\mathcal{A})$ respectively as the "reward" and "cost" functions, and ρ as the "budget" that one is allowed to spend. Next we discuss its optimization.

6.2 Properties of Objective Function

It is noticed that the objective function $R(\mathcal{A})$ in Eqn.(6) possesses a set of interesting properties.

- Non-negative - $R(\emptyset) = 0$; that is, k's creativity does not change if no reading paper is taken into account.
- Non-decreasing - If $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{S}$, then $R(\mathcal{A}) \leq R(\mathcal{B})$. Intuitively, reading more papers can only reduce k's creativity.
- Diminishing return - Given $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{S}$, reading one more paper in addition to $\mathcal{A}$ improves $R(\cdot)$ at least as much as reading it in addition to $\mathcal{B}$. More formally,

Theorem 1. *Given $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{S}$ and a paper $x^* \in \mathcal{S} \setminus \mathcal{B}$, it holds that*

$$R(\mathcal{A} \cup \{x^*\}) - R(\mathcal{A}) \geq R(\mathcal{B} \cup \{x^*\}) - R(\mathcal{B})$$

Thus, $R(\cdot)$ is a submodular function [20].

Proof. First note that in $R(\cdot)$, for a given target paper k, $\ell(\{\varphi_{i,j}\}_{i,j \in \mathcal{C}_k})$ is input-independent. We thus focus our discussion on $\ell(\{\varphi_{i,j} - \Delta_{\cdot}^{i,j}\}_{i,j \in \mathcal{C}_k})$. Let us consider $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{S}$ and a reading paper $x^* \in \mathcal{S} \setminus \mathcal{B}$.

For ease of exposition, we assume that $\mathcal{A}$ (and $\mathcal{B}$) contains a special element $\oslash$, which corresponds to the case that no paper is selected from $\mathcal{A}$ (and $\mathcal{B}$). For each pair $i, j \in \mathcal{C}_k$, $\Delta_{\mathcal{A}}^{i,j} = \max_{x \in \mathcal{A}} \Delta_x^{i,j}$ and $\Delta_{\mathcal{B}}^{i,j} = \max_{x \in \mathcal{B}} \Delta_x^{i,j}$. Since $\mathcal{A} \subseteq \mathcal{B}$, it holds that $\Delta_{\mathcal{A}}^{i,j} \leq \Delta_{\mathcal{B}}^{i,j}$. Let $x' = \arg\max_{x \in \mathcal{B} \cup \{x^*\}} \Delta_x^{i,j}$. We consider the following three cases.

(i) $x^* = x'$. It is clear that $\Delta_{\mathcal{A} \cup \{x^*\}}^{i,j} = \Delta_{\mathcal{B} \cup \{x^*\}}^{i,j} = \Delta_{x^*}^{i,j}$; thus $\Delta_{\mathcal{A} \cup \{x^*\}}^{i,j} - \Delta_{\mathcal{A}}^{i,j} \geq \Delta_{\mathcal{B} \cup \{x^*\}}^{i,j} - \Delta_{\mathcal{B}}^{i,j}$.

(ii) $x^* \neq x'$ and $x' \in \mathcal{A}$. This implies $\Delta_{\mathcal{B} \cup \{x^*\}}^{i,j} = \Delta_{\mathcal{A} \cup \{x^*\}}^{i,j} = \Delta_{\mathcal{B}}^{i,j} = \Delta_{\mathcal{A}}^{i,j}$. Thus, $\Delta_{\mathcal{A} \cup \{x^*\}}^{i,j} - \Delta_{\mathcal{A}}^{i,j} = \Delta_{\mathcal{B} \cup \{x^*\}}^{i,j} - \Delta_{\mathcal{B}}^{i,j}$.

(iii) $x^* \neq x'$ and $x' \notin \mathcal{A}$. This implies $\Delta_{\mathcal{B} \cup \{x^*\}}^{i,j} = \Delta_{\mathcal{B}}^{i,j}$ and $\Delta_{\mathcal{A} \cup \{x^*\}}^{i,j} \geq \Delta_{\mathcal{A}}^{i,j}$, i.e., $\Delta_{\mathcal{A} \cup \{x^*\}}^{i,j} - \Delta_{\mathcal{A}}^{i,j} \geq \Delta_{\mathcal{B} \cup \{x^*\}}^{i,j} - \Delta_{\mathcal{B}}^{i,j}$.

In all three cases above, it holds that $\Delta_{\mathcal{A} \cup \{x^*\}}^{i,j} - \Delta_{\mathcal{A}}^{i,j} \geq \Delta_{\mathcal{B} \cup \{x^*\}}^{i,j} - \Delta_{\mathcal{B}}^{i,j}$. Because the aggregation function $\ell(\cdot)$ (e.g., average, percentile, maximum) is non-decreasing (cf § 4), it leads to $R(\mathcal{A} \cup \{x^*\}) - R(\mathcal{A}) \geq R(\mathcal{B} \cup \{x^*\}) - R(\mathcal{B})$. □

6.3 Optimization Algorithm

In general, maximizing a submodular function is known to be NP-hard [20]. Yet, as each paper $x \in \mathcal{S}$ has unit cost, a greedy algorithm is applicable, which provides near-optimal guarantee for the found results. Next we introduce such an algorithm, as sketched in Algorithm 2.

Let $\mathcal{A}_s$ denote the set of selected papers after the s-th step. Starting with an empty set $\mathcal{A}_0 = \emptyset$, and iteratively, at the s-th step, it adds the paper $x^* \in \mathcal{S}$ to $\mathcal{A}_{s-1}$, such that the following marginal gain is maximized:

$$x^* = \arg\max_{x \in \mathcal{S} \setminus \mathcal{A}_{s-1}} R(\mathcal{A}_{s-1} \cup \{x\}) - R(\mathcal{A}_{s-1})$$

This process stops if the marginal gain reaches zero or the budget is used up.

Algorithm 2 provides the following near-optimal guarantee for the found enablers.

Theorem 2 ([20]). *Given that the function $R(\cdot)$ is submodular, nondecreasing, and $R(\emptyset) = 0$, then the greedy algorithm finds $\mathcal{A}$, such that $R(\mathcal{A}) \geq (1 - 1/e) \max_{|\mathcal{A}'| = \rho} R(\mathcal{A}')$.*

Algorithm 2: Finding minimum set of enablers

```
   Input: target paper $k$, existing literature $\mathcal{S}$, budget $\rho$
   Output: minimum set of enablers $\mathcal{A}$
1  $\mathcal{A} \leftarrow \emptyset$;
2  for $s = 1, \ldots, \rho$ do
       /* greedy approach to find the next enabler */
3      for each $x \in \mathcal{S} \setminus \mathcal{A}$ do
4          compute $R(\mathcal{A} \cup \{x\}) - R(\mathcal{A})$;
5      while true do
6          find $x^* = \arg\max_{x \in \mathcal{S} \setminus \mathcal{A}} R(\mathcal{A} \cup \{x\}) - R(\mathcal{A})$;
           /* temporal dynamics of information production-consumption dependency */
7          pick $x^*$ with probability $m(\tau^p_{x^*,k})$;
8          if $x^*$ is picked then break else remove $x^*$ from $\mathcal{S}$
9      if $R(\mathcal{A} \cup \{x^*\}) - R(\mathcal{A}) = 0$ then break $\mathcal{A} = \mathcal{A} \cup \{x^*\}$;
10 return $\mathcal{A}$;
```

6.4 Extensions

Next we extend Algorithm 2 along two directions: (i) handling the case of multiple target papers and (ii) taking account of temporal dynamics of information production-consumption dependency.

Multiple Target Papers

Let $R(k, \mathcal{A}) = \ell(\{\varphi_{i,j}\}_{i,j \in \mathcal{C}_k}) - \ell(\{\varphi_{i,j} - \Delta^{i,j}_{\mathcal{A}}\}_{i,j \in \mathcal{C}_k})$. We update the objective function as:

$$R(\mathcal{A}) = \sum_{k \in \mathcal{P}} \lambda_k R(k, \mathcal{A})$$

where λ_k is paper k's weight, indicating its importance, with $\forall k \in \mathcal{P}, \lambda_k \geq 0$ and $\sum_{k \in \mathcal{P}} \lambda_k = 1$.

Under this multi-criterion setting, there may be cases that two solutions $\mathcal{A}, \mathcal{A}'$ are incompatible, i.e., $R(k, \mathcal{A}) > R(k, \mathcal{A}')$ while $R(k', \mathcal{A}) < R(k', \mathcal{A}')$. Instead of looking for the optimal solution, we resort to finding a Pareto-optimal solution [2]. A solution $\mathcal{A}$ is Pareto-optimal if no other solution $\mathcal{A}'$ satisfies that (i) $R(k, \mathcal{A}') \geq R(k, \mathcal{A})$ for all $k \in \mathcal{P}$ and $k \neq k'$, and (ii) $R(k', \mathcal{A}') > R(k', \mathcal{A})$ for a specific $k' \in \mathcal{P}$.

As $R(\cdot)$ is a non-negative linear combination of submodular functions, any solution that maximizes $R(\cdot)$ is guaranteed to be Pareto-optimal [2]. Moreover, since the submodularity is closed under the non-negative linear combinations, Algorithm 2 can be readily applied.

Temporal Dynamics

Furthermore, we take into account the temporal dynamics of information production-consumption dependency (cf. § 5.2). To build the temporal dynamics model, for $x \in \mathcal{C}_k \cap \mathcal{Q}$, we examine x's publishing time (t^p_x) and k's publishing time (t^p_k). The difference ($\tau^p_{x,k} = t^p_k - t^p_x$) reflects the temporal gap between x and k's publishing. Figure 8(b) shows that this interval fits a log-log distribution, denoted by $Pr(\tau)$. By discretizing and normalizing $Pr(\tau)$, we compute the probability that a paper published Δt years ago influences a current paper, i.e., $m(\Delta t) = \int_{\tau=\Delta t}^{+\infty} Pr(\tau) \mathrm{d}\tau$. Similar to Algorithm 1, we incorporate this temporal dynamics into Algorithm 2 (line 7-9). We omit the details due to space limitations.

6.5 Empirical Study

To validate the effectiveness of our solution, we apply Algorithm 2 to predicting the most influential enablers for papers published by Indiana University in a specific year t. Specifically, we assume that each paper

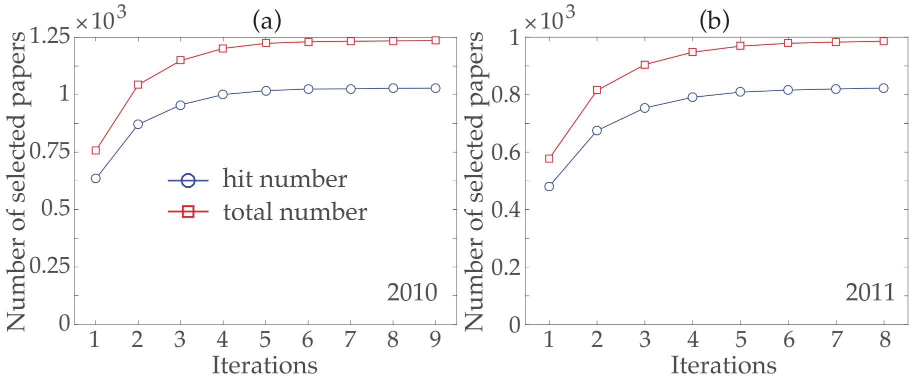

Figure 11: Effectiveness of predicting the most influential enablers for the papers published in (a) 2010 and (b) 2011.

$k \in \mathcal{P}_t$ is of uniform importance, i.e., $\lambda_k = 1/|\mathcal{P}|$. We consider all papers published before year t in the Mag dataset as the literature space $\mathcal{S}$ (about 20 million papers for 2010 and 2011). We then measure the precision of our prediction algorithm: the proportion of selected papers $\mathcal{A}$ that appear in the reading set $\mathcal{Q}$ of the Click dataset, i.e., $|\mathcal{A} \cap \mathcal{Q}|/|\mathcal{A}|$.

Figure 11 illustrates the measurement results for $t = 2010$ and 2011. It is observed that our prediction model is fairly scalable: in both cases, it quickly converges using less than 10 iterations. Furthermore, among the selected papers $\mathcal{A}$, most of them indeed appear in the reading set $\mathcal{Q}$ (with precision over 83% in both cases), highlighting the effectiveness of our predictive model.

7 Conclusion & Discussion

In this work, we conducted a systematic study on creativity in scientific enterprise. For the first time, by directly correlating authors' raw information consumption behaviors with their knowledge products, we found remarkable predictability in scientific creative processes: of over 59.0% papers across all scientific fields, 25.7% of their creativity can be readily explained by information consumed by their authors. Leveraging these findings, we proposed a predictive framework that captures the impact of authors' information consumption over their future knowledge products. By using two web-scale, longitudinal real datasets, we demonstrated the efficacy of our framework in identifying the most critical knowledge to fostering target scientific innovations. Our framework is not limited to scientific creative processes only. Indeed, its mechanistic nature makes it potentially applicable for describing creative processes in other domains as well, such as musical, artistic and linguistics creativity.

This work also opens up several directions that are worth future investigations. For example, due to privacy and technology constraints, our study tracks information consumption and knowledge production at an organizational level. Thus, extending such study to an individual level could be fruitful and potentially shed new light on the nature of creativity. Furthermore, recent work has shown that various semantic features (e.g., author, content, venue) can be used to predict long-term impact of scientific artifacts in their early stages. Such semantic features could be integrated into our framework to train microscopic (author-, content-, and venue-specific) creativity models. Lastly, our model makes falsifiable prediction for creative processes, making it a viable candidate to assess and guide experimental studies, results of which can feed back to and improve the model with more accurate and realistic predictions.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, 1994.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, New York, NY, USA, 2004.

[3] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, 2014.

[4] R. Collins. *The Sociology of Philosophies: A Global Theory of Intellectual Change.* Belknap Press of Harvard University Press, 1998.

[5] S. Colton. Creativity versus the perception of creativity in computational systems. In *AAAI Spring Symposium: Creative Intelligent Systems'08*, 2008.

[6] M. Csikszentmihalyi. *Creativity-flow and the psychology of discovery and invention.* Harper perennial, 1996.

[7] T. De Smedt. Modeling Creativity: Case Studies in Python. *ArXiv e-prints*, 2014.

[8] S. Doboli, F. Zhao, and A. Doboli. New measures for evaluating creativity in scientific publications. *ArXiv e-prints*, 2014.

[9] Y. Dong, R. A. Johnson, and N. V. Chawla. Will this paper increase your h-index?: Scientific impact prediction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, 2015.

[10] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW (Physics).* Oxford University Press, Inc., New York, NY, USA, 2003.

[11] J. A. Evans and J. G. Foster. Metaknowledge. *Science*, 331(6018):721–725, 2011.

[12] L. Fleming. Recombinant uncertainty in technological search. *Management Science*, 47(1):117–132, 2001.

[13] L. Gabora and A. Saab. Creative interference and states of potentiality in analogy problem solving. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, COGSCI '13, 2013.

[14] M. R. Guevara, D. Hartmann, M. Aristarán, M. Mendoza, and C. A. Hidalgo. The Research Space: using the career paths of scholars to predict the evolution of the research output of individuals, institutions, and nations. *ArXiv e-prints*, 2016.

[15] B. F. Jones, S. Wuchty, and B. Uzzi. Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, 322(5905):1259–1262, 2008.

[16] D. Kim, D. Burkhardt Cerigo, H. Jeong, and H. Youn. Technological novelty profile and invention's future impact. *ArXiv e-prints*, 2015.

[17] A. Koestler. *The Act of Creation.* Arkana, 1964.

[18] L. Li and H. Tong. The child is father of the man: Foresee the success at the early stage. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 2015.

[19] M. Meiss, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. Ranking web sites with real user traffic. In *Proc. First ACM International Conference on Web Search and Data Mining (WSDM)*, 2008.

[20] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978.

[21] Plato. *The republic, Book X.*

[22] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[23] R. Saunders and J. S. Gero. Artificial creativity: A synthetic approach to the study of creative behaviour. In *Computational and Cognitive Models of Creative Design V*, 2001.

[24] W. Shadish and S. Fuller. *The social psychology of science*. Guilford Press, 1994.

[25] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, 2015.

[26] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

[27] T. Veale and Y. Hao. Learning to understand figurative language: From similes to metaphors to irony. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, COGSCI '07, 2007.

[28] D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[29] M. Weitzman. Recombinant growth. *Quarterly Journal of Economics*, 113(2):331–360, 1998.

[30] G. A. Wiggins. A preliminary framework for description, analysis and comparison of creative systems. *Know.-Based Syst.*, 19(7):449–458, 2006.