Serendipity or Preparedness? Quantifying Creativity in Scientific Enterprise

ABSTRACT

Human creativity is the ultimate driving force behind all of scientific progress. The building blocks of innovations are often embodied in existing knowledge, yet it is creativity that blends seemingly disparate concepts, ideas and theories. Existing studies attempt to understand this phenomenon by investigating reference relationships of scientific publications. Unfortunately, an article's references, at best, only partially or remotely reflect its authors' actual information consumption, highlighting a fundamental reality-perception gap in current effort. In this paper, using two Web-scale longitudinal datasets (120.1 million papers across all scientific fields and 53.5 billion Web requests spanning over 4.5 years), we explore the predictability in scientific creativity. By directly correlating authors' *raw* information consumption behaviors with their publications, we find remarkable reproducible patterns [#] in scientific creative processes across all scientific fields. Further, by leveraging these findings, we develop SERENDIP, a first-of-its-kind mechanistic framework, that not only effectively predict disparate references likely to be connected by creativity, but also accurately identifies critical references for such linkings to happen. We believe our framework is of fundamental importance to studies of scientific creativity. It promotes strategies to foster creative processes and provides insights towards a more effective design of information recommendation platforms.

1. INTRODUCTION

> *It is the function of creative man to perceive and to connect the seemingly unconnected.*
>
> William Plomer

Of the many propulsions behind scientific progress, one stands out in its tremendous yet intangible force: human creativity [15, 22]. Scientific innovations are often prompted through effective combinations of existing knowledge [29, 13, 10], yet it is creativity that spurs and catalyzes the process of connecting seemingly disparate concepts, ideas, and theories. A plethora of experimental studies in psychology and cognitive science have offered various theries for the phenomena of creativity [5, 12]. Despite the significance of these studies, there is still a lack of quantitative understanding of human creativity, not to mention mechanistic models to describe such processes.

Nowadays, thanks to prolific scientific publication archives and citation indices (e.g., DBLP[1], PUBMED[2], WEB OF SCIENCE[3]), we are equipped with the lens to study creativity in scientific research with unprecedented granularity and precision. Existing studies focus on quantitatively measuring the creativity of scientific artifacts (e.g., papers and patents) by investigating their reference relationships [25, 7, 14]. While these studies offer convincing evidences that creativity can be modeled as intersections of originally disconnected references, little is know how such intersections are triggered.

In this work, we directly address this gap. We argue that, to understand scientific creative processes, solely relying on publications' reference relationships is insufficient. An article's references, at best, only partially or remotely reflect its authors' actual information intake. First, the references may not include the most relevant literature (e.g., the authors fail to identify the most related article among multiple similar ones). Further, the references may not provide a comprehensive view of all literature inspiring the authors (e.g., due to space limitations). Finally, to understand the correspondence between information consumption and production behaviors, it is imperative to characterize their temporal correlation; however, the references alone do not indicate when the cited literature is actually read by the authors. All these drawbacks highlight a fundamental gap between reality and perception in current effort.

Thus, in this work, we directly contrast authors' *raw* information intake with their publications. Specifically, using two large longitudinal datasets, Microsoft Academic Graph Dataset (120.1 million publications across all scientific fields) and Indiana University Click Dataset (53.5 billion web requests by 100K users spanning over 4.5 years), we explore predictable patterns in creative processes of scientific publications. Although varied privacy and technology constraints preclude the possibility of tracking information consumption and production at an individual level, by studying such correlations at an organization level, we find that creativity in scientific publications follows remarkably reproducible patterns ([fill # here] predictability) across about [fill # here] authors and all research fields.

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

Leveraging these insights, we develop a mechanism framework, SERENDIP, to model the impact of authors' information intake over their publications. Using the above datasets as an exemplary case, we demonstrate that SERENDIP not only effectively predict disparate references likely to be connected by creativity, but also accurately identifies critical references necessary for such linkings to happen. As a result, SERENDIP can make valuable information recommendation to authors' creative processes.

To our best knowledge, this work represents one of the first few quantitative frameworks to model creative processes of scientific research within the context of authors' information consumption behaviors. We believe our method is of fundamental importance to studies of human creativity, promotes new strategies in stimulating creative thinking, and provides significant insights towards a design of effective information recommendation platforms.

The remainder of the paper proceeds as follows. Section 2 surveys relevant literature. Section 3 describes the datasets used in our study. Section 4 presents our empirical study on the predictability in scientific creative processes. Section 5 details the model design of SERENDIP and develops efficient inference algorithms to fit the model. Section 6 discusses the application of SERENDIP in predicting future innovations and recommending critical references. Section 7 empirically evaluates the proposed models and algorithms. The paper is concluded in Section 8.

2. RELATED WORK

In this section, we review three categories of related work, namely, assessment of creativity, scientific impact prediction, and computational creativity.

Scholarly interest in creativity has spurred empirical studies pertaining to various disciplines (e.g., psychology, cognitive science, economics, and philosophy [15, 5, 3, 29, 11, 12]). Despite the significance of these studies, there is still a lack of quantitative understanding of creativity, not to mention mechanistic models. Now, thanks to the proliferation of scientific publication archives and citation indices, it is feasible to study creativity with unprecedented granularity and precision. One active line of enquiry is to develop meaningful measurment of scientific creativity. For example, Uzzi et al. [25] measured a scentific article's creativity as atypical pairwise combinations of references in its bibliography; Fleming [11] measured a patent's innovation using new combinations of patents cited in its references. Despite their varied concrete forms, this line of research shows convincingly that in general, creativity can be modeled as intersections between originally disconnected knowledge. However, little is known hitherto how these intersections are actually triggered. Our work directly addresses this gap. To the best of our knowledge, this work is among the first few to quantitatively model creative processes by taking raw information consumption behaviors into account.

Meanwhile, a growing body of work has focused on predicting long-term impact (primarily measured by citations) of scientific artifacts during their early stages [27, 2, 8, 17] using various semantic features (e.g., author, content, venue). While they are useful for projecting the full citation trajectories of scientific publications (i.e., one way of measuring their novelty), insights extracted from semantic features are incapable of explaining how such novelty is established. However, these studies are complementary to our work, in a sense that the useful semantic features learned can be integrated into our framework to train microscopic (e.g., author-, content-, and venue-specific) creativity models.

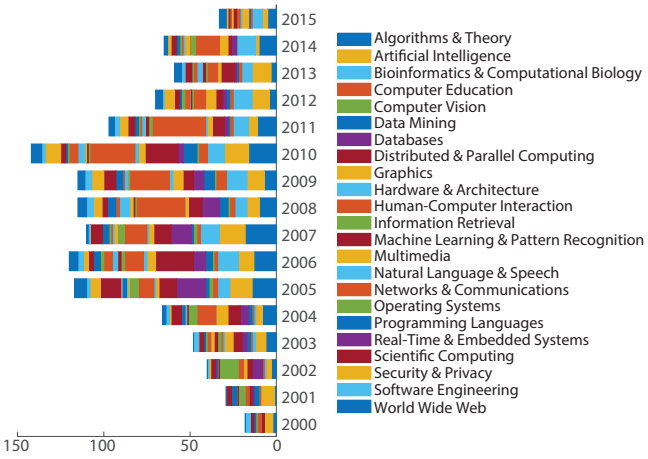

Figure 1: Number of publications in each sub-area of computer science per year.

Finally, our work is related to the broad area of computational creativity [30, 6], which focuses on developing artificial intelligence models that exhibit and generate creativity, such as problem solving [21], visual creativity [4], and linguistic creativity [26]. In contrast, our work focuses on understanding and modeling creative processes in scientific publications. However, incorporating such intelligence models to enhance the predictive power of SERENDIP would be one promising future direction.

This work also draws connections to other modeling effort. The design of SERENDIP model is inspired by the social network evolution phenomena and the link prediction models [16, 28, 20, 9] in particular. However, our work differs in that (i) our focus is to understand the impact of adding new edges to certain network properties and (ii) our criterion function of selecting new edges to be added is completely different from the prior work.

3. DATA

We start with describing in detail the datasets used in our empirical study.

3.1 Raw Data

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

At present, the most comprehensive data that captures information consumption of scientific research is perhaps web traffic generated by researchers, reflecting how they request and access online resources (e.g., publication archives). Therefore, in our study, we used the Indiana University Click Dataset [19] (CLICK), which records about 53.5 billion web requests initiated by users at Indiana University over the period from September 2006 to May 2010. This anonymized dataset was collected by applying a Berkeley Packet Filter to web traffic passing through the border router of Indiana University and matching all the traffic containing Http GET requests. Each record consists of the following fields: ⟨timestamp, requested url, referring url, agent, flag⟩, in which "agent" indicates whether the user agent was a browser or

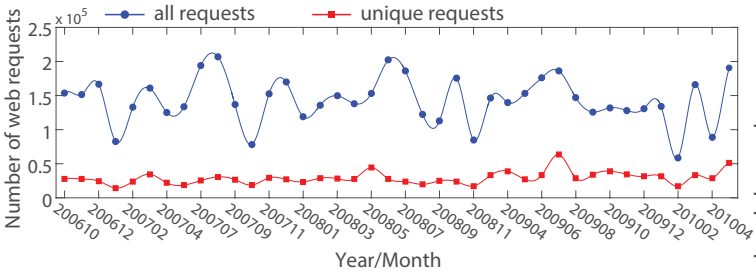

Figure 2: Total number of requests and number of unique requests per month.

a bot while "flag" indicates whether the request was generated inside or outside of Indiana University. We filtered all incoming, bot-generated requests.

Meanwhile, to capture information production of scientific research, we used the Microsoft Academic Graph Dataset [23] (MAG). In a nutshell, MAG is a Web-scale entity graph comprising scientific publication records, reference relationships between publications, as well as authors, affiliations, venues (i.e., journals and conference) and fields of study (i.e., subjects). As of November 6, 2015, the MAG dataset constitutes 120.9 million articles published over 24,843 venues across all scientific fields.

Furthermore, we also queries the service of Microsoft Academic Search (MAS)[4] to collect metadata of the venues in the MAG dataset, including their subjects, total number of publications, and total number of citations.

3.2 Preprocessing

To correlate the aforementioned two datasets, in the MAG dataset, we identified all the publications that have at least one author affiliated with Indiana University and were published during the period from 2007 to present, with a total of 24,399 publications. More specifically, Figure 1 illustrates the number of publications in each subject of computer science from 2007 to 2015. Note that both the number and constitution of publications vary significantly from year to year, suggesting a lack of order and hence lack of predictability in creative processes. However, as we will show next, this lack of predictability is only apparent, because creative processes follow highly reproducible dynamical patterns once their information input is taken into account.

Meanwhile, in the CLICK dataset, we identified all the web traffic that requested for publications in the MAG dataset by matching URLs embedded in the requests against URLs of the publications in online archives. The resulted dataset consists of 5.8 million records requesting for 4.6 million unique publications (e.g., unique requests). Figure 2 illustrates the total number of requests and the number of unique ones per month from September 2006 to May 2010. It is noted that while the total number of requests fluctuate significantly, the monthly unique requests remain fairly stable across the time period, implying constant amount of information input.

4. PREDICTABILITY IN CREATIVITY

This section presents our empirical study on predictability in scientific creative processes. For ease of presentation, we first introduce a set of fundamental concepts and assumptions used throughout the paper.

[4]MAS: http://academic.research.microsoft.com

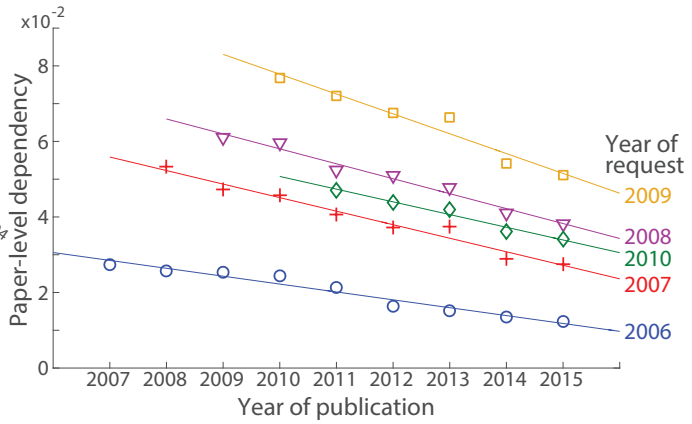

Figure 3: Illustrations of paper-level dependency of published papers on requested web resources.

4.1 Preliminaries

We refer to a scientific publication as a "paper", denoted by p. We assume that p is described by a tuple $\langle t_p, v_p, \mathcal{K}_p, \mathcal{R}_p \rangle$, which represent the publication time, the publication venue, the keyword(s), and the references of p, respectively. Further, we refer to each access (e.g., browsing or download) of online publication archives as a "request", denoted by q. We assume that q is described by a tuple $\langle t_q, p_q \rangle$, representing the time of access and the requested paper.

Let $\mathcal{P}_{t,t'}$ denote the set of papers published during the time period from t to t'. In particular, $\mathcal{P}_{,t}$ represents all the papers published till t. Similarly, let $\mathcal{Q}_{t,t'}$ denote the set of requests made within the time window of $[t, t']$. When the context is clear, we use $\mathcal{Q}_{t,t'}$ to denote the collection of papers requested within this time window as well.

Given the facts that (i) for a number of venues in the MAG dataset, only their publication year is specified and (ii) even with more granular timestamps, the publication time of a paper typically does not precisely reflect when it is actually finished, we use year as the default time granularity in our study, unless noted otherwise. Therefore, with a little abuse of notations, we may use t to denote a timestamp or a one-year-long time window. As an example, $\mathcal{P}_t$ may represent the set of published papers in year t.

4.2 Dependency of Publications on Requests

Next, we explore predictability in scientific creative processes by empirically measuring the correlation between requests and publications. Specifically, for a given set of publications $\mathcal{P}_t$, we compare the collection of prior work referenced by $\mathcal{P}_t$ against the set of requests $\mathcal{Q}_{t'}$ at an early time t' ($t' < t$). More formally, let $\mathcal{R}(\mathcal{P}_t)$ be the set of prior work referenced by papers in $\mathcal{P}_t$:

$$\mathcal{R}(\mathcal{P}_t) = \cup_{p \in \mathcal{P}_t} \mathcal{R}_p$$

We measure the dependency of publications $\mathcal{P}_t$ on requests $\mathcal{Q}_{t'}$ at two different levels.

Paper-Level Dependency

We start with directly comparing reference set $\mathcal{R}(\mathcal{P}_t)$ against request set $\mathcal{Q}_{t'}$. In specific, we measure paper-level dependency of $\mathcal{P}_t$ on $\mathcal{Q}_{t'}$, denoted by $\Phi_{\text{P}}(\mathcal{P}_t, \mathcal{Q}_{t'})$, by computing

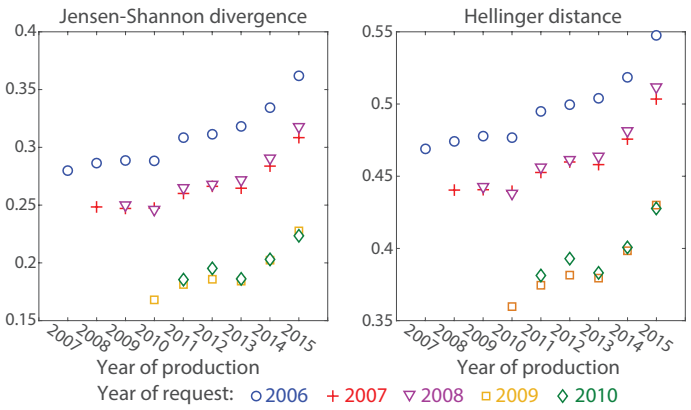

Figure 4: Illustrations of topic-level dependency of published papers on requested web resources.

the Jaccard's coefficient of $\mathcal{R}(\mathcal{P}_t)$ and $\mathcal{Q}_{t'}$:

$$\Phi_{\mathrm{P}}(\mathcal{P}_t, \mathcal{Q}_{t'}) = \frac{|\mathcal{R}(\mathcal{P}_t) \cap \mathcal{Q}_{t'}|}{\min\{|\mathcal{R}(\mathcal{P}_t)|, |\mathcal{Q}_{t'}|\}}$$

Figure 3 illustrates the results of our measurement of the paper-level dependency in our datasets. Specifically, for each t varying from 2007 to 2015, we measure $\Phi_{\mathrm{P}}(\mathcal{P}_t, \mathcal{Q}_{t'})$ for t' ranging from 2006 to $(t-1)$.

It is observed that the dependency of future publications on input information demonstrates interesting temporal dynamics. In particular, we have the following key observations. First, for a given set of requests $\mathcal{Q}_{t'}$ (i.e., fixed t'), the dependency of future publications $\mathcal{P}_t$ on $\mathcal{Q}_{t'}$ decreases as t grows, implying that the influence of $\mathcal{Q}_{t'}$ gradually decays over time. Second, for a given set of publications $\mathcal{P}_t$ (i.e., fixed t), their dependency on $\mathcal{Q}_{t'}$ increases with t', implying that more recent information consumption carries more weight in future publications. The only exception is the case corresponding to requests made in 2010, which is explained by the that the CLICK dataset only includes web traffic over January $\sim$ May 2010.

Topic-Level Dependency

We then extend the study of the association of publication set $\mathcal{P}_t$ and request set $\mathcal{Q}_{t'}$ to the topic level.

As aforementioned, each keyword in the MAG dataset is mapped to a topic from a finite set of topics $\mathcal{T}$. Thus one is able to map each paper to one or more topics. 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*".

We encode the topics of paper p using a $|\mathcal{T}|$-dimensional vector $\boldsymbol{t}_p$, with the i-th element to be 1 if p is associated with the i-th topic and "0" otherwise. We aggregate such topic vectors of all the papers in $\mathcal{R}(\mathcal{P}_t)$ and $\mathcal{Q}_{t'}$ as $\boldsymbol{t}(\mathcal{P}_t)$ and $\boldsymbol{t}(\mathcal{Q}_{t'})$, respectively. Formally,

$$\boldsymbol{t}(\mathcal{P}_t) = \sum_{p \in \mathcal{R}(\mathcal{P}_t)} \boldsymbol{t}_p \qquad \boldsymbol{t}(\mathcal{Q}_{t'}) = \sum_{p \in \mathcal{Q}_{t'}} \boldsymbol{t}_p$$

We then normalize both topic vectors to ensure that the elements across all the dimensions sum up to 1. Let $\bar{\boldsymbol{t}}(\mathcal{P}_t)$ and $\bar{\boldsymbol{t}}(\mathcal{Q}_{t'})$ denote the normalized topic vectors, which can be considered as distributions.

We now measure the topic-level dependency of $\mathcal{P}_t$ on $\mathcal{Q}_{t'}$, $\Phi_{\mathrm{T}}(\mathcal{P}_t, \mathcal{Q}_{t'})$ using two distribution-similarity metrics.

The first one is the Jensen-Shannon divergence:

$$\Phi_{\mathrm{T}}(\mathcal{P}_t, \mathcal{Q}_{t'}) = \frac{1}{2}\mathtt{D}_{\mathtt{KL}}(\bar{\boldsymbol{t}}(\mathcal{P}_t)||\bar{\boldsymbol{t}}(\mathcal{Q}_{t'})) + \frac{1}{2}\mathtt{D}_{\mathtt{KL}}(\bar{\boldsymbol{t}}(\mathcal{Q}_{t'})||\bar{\boldsymbol{t}}(\mathcal{P}_t))$$

where $\mathtt{D}_{\mathtt{KL}}(\cdot||\cdot)$ represents the Kullback-Leibler divergence of two distributions.

The second is the Hellinger distance, given as

$$\Phi_{\mathrm{T}}(\mathcal{P}_t, \mathcal{Q}_{t'}) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{|\mathcal{T}|}\left(\sqrt{\bar{\boldsymbol{t}}_i(\mathcal{P}_t)} - \sqrt{\bar{\boldsymbol{t}}_i(\mathcal{Q}_{t'})}\right)^2}$$

where $\bar{\boldsymbol{t}}_i(\cdot)$ is the i-th dimension of the topic vector.

For each pair of t and t', we measure $\Phi_{\mathrm{T}}(\mathcal{P}_t, \mathcal{Q}_{t'})$ in terms of both Jensen-Shannon divergence and Hellinger distance, with the results illustrated in Figure 4.

We have the following key observations. First, the topic-level dependency demonstrates pattens similar to the paper-level dependency: (i) the influence of information consumption over future publications decays over time; and (ii) more recently consumed information exerts stronger influence over future publications. Second, compared with the paper-level dependency, the topic-level dependency seems less "stratified" in that adjacent years show more similar topic distributions. For example, in both plots of Figure 4, the dependencies with respect to requests made in 2007 and 2008 show strong resemblance. This phenomenon may be explained by that researchers' interests in research topics are more stable than their interests in concrete papers.

4.3 Explanation of Creativity

The study of paper-level and topic-level dependencies of future publications over information consumption confirms our conjecture that authors' future publications are heavily influenced by the prior work they are currently reading. This also hints that it is conceivable to answer the following key question:

Can the information currently consumed by the authors help explain, to a certain extent, the creativity demonstrated in their future publications?

To answer this question, the foremost question we need to address is how to assess creativity quantitatively. Existing studies across different disciplines (e.g., [11, 25]) offer convincing evidences that the creativity of a paper can be effectively measured by the disparity of the prior which it is built upon. Intuitively, a paper that connects otherwise remotely unrelated prior work is conceived to be novel. Following this paradigm, we introduce a general network-centric creativity model, which subsumes a range of existing ones [11, 25].

Network-Centric Creativity Model

Given time t, we construct a reference network $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ for all papers (not only from the target organization) published till t, where papers $u, v \in \mathcal{V}_t$ are adjacent, i.e., $\overline{uv} \in \mathcal{E}_t$, only if u (or v) cites v (or u). Note that for simplicity, we use an undirected network model, however, it is straightforward to extend to directed networks to account for the forward/backward directions of references. $\mathcal{G}_t$ can be augmented by incorporating auxiliary information. In this work, we particularly specify a labeling function $\mathcal{L} : \mathcal{V} \to \mathcal{T}$, which maps each paper to its topic(s). Next we use $\mathcal{G}_t$ as the frame of reference to evaluate the creativity of a given paper. Note that $\mathcal{G}_t$ evolves over time.

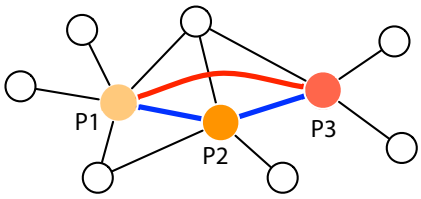

Figure 5: Connection $P_1 \sim P_3$ introduced by exposure to disparate prior work P_1, P_2 and P_3.

Now, for paper p published at time t, we examine its referenced prior work $\mathcal{R}_p$. For each pair of prior work $x, y \in \mathcal{R}_p$, we measure their network proximity in $\mathcal{G}_t$ to captures their relevance.

A plethora of network proximity measures have been proposed in literature, including Common Neighbors, Jaccard, Preferential Attachment, Adamic-Adar, and Katz (refer to [18] for a survey). In our implementation, we extend the Random Walk with Restart (RWR) [24], which performs extremely well in our experimental study. Specifically, consider a random walker starting from paper x, who will iteratively moves to a random neighboring paper in $\mathcal{G}_t$ with probability $c\beta$, teleport to a random paper of the same topic with probability of $c(1-\beta)$, and returns to x with probability $(1-c)$. This model essentially captures relevance of papers reflected in reference relationships and topic similarity.

Denote by $\mathsf{q}_{x,y}$ the stationary probability over y with a walk starting from x. Then the novelty of the combination of x and y is given by:

$$1 - \frac{1}{2}\mathsf{q}_{x,y} - \frac{1}{2}\mathsf{q}_{y,x}$$

Intuitively, if x and y are more loosely connected, we consider their combination to be more novel.

The novelty of paper p is the aggregation of novelty of all possible pairs of its referenced prior work:

$$\Psi_p = 1 - \frac{\sum_{x,y \in \mathcal{R}_p} (\mathsf{q}_{x,y} + \mathsf{q}_{y,x})}{2\binom{|\mathcal{R}_p|}{2}} \quad (1)$$

[plot distributions of shortest path length and rwr similarity].

Serendipity + Preparedness = Creativitiy

Next we intend to model the influence of information consumption over creative processes. At a high level, the exposure to disparate literature offers the opportunity of making connections between disparate prior work. As an example, consider the following three papers:

- P_1: "*Fast Algorithms for Mining Association Rules*". R. Agrawal and R. Srikant. VLDB '94.
- P_2: "*CloseGraph: Mining Closed Frequent Graph Patterns*". X. Yan and J. Han. KDD '03.
- P_3: "*Graph Indexing: A Frequent Structure-based Approach*". X. Yan, P. Yu and J. Han. SIGMOD '04.

where P_3 cites P_2 (but not P_1) while P_2 cites P_1. Although P_3 differs from P_1 significantly, after reading all three papers, one is able to draw the connection from P_1 to P_3 as "*frequent pattern mining*" $\rightarrow$ "*frequent graph pattern mining*" $\rightarrow$ "*frequent graph pattern based indexing*". This process is illustrated in Figure 5.

To capture the intuition that literature reading helps make connections between prior work, we model the influence of information consumption as *edge addition* operations to existing reference network $\mathcal{G}_t$. For instance, the exposure to all P_1, P_2 and P_3 introduces an additional edge $\overline{P_1P_3}$ to the reference network.

Denote by $\mathcal{G}'_t$ the reference network after such edge addition operations. Clearly, the addition of extra edges "shortcut" some paths in $\mathcal{G}_t$; therefore, we tend to observe lower creativity measure of p with respect to $\mathcal{G}'_t$. Let $\Psi(p|\mathcal{G}_t)$ and $\Psi(p|\mathcal{G}'_t)$ denote the creativity measures of p over $\mathcal{G}_t$ and $\mathcal{G}'_t$ respectively. The change of creativity measure, given as:

$$\Delta(p|\mathcal{G}_t, \mathcal{G}'_t) = \Psi(p|\mathcal{G}_t) - \Psi(p|\mathcal{G}'_t)$$

captures the part of creativity that can be explained by the authors' information consumption, which we refer to as the part of "preparedness". Meanwhile, the creativity remaining in $\mathcal{G}'_t$ cannot be explained by the information consumed by the authors, which we refer to as the part of "serendipity". Roughly speaking, the creativity of a paper reflects an superimpose of both effects, i.e.,

$$\text{Serendipity} + \text{Preparedness} = \text{Creativitiy}$$

Despite its attractive simplicity, to use this theory in practice, we need a mechanistic framework to implement the theory to explain real datasets. Next we present SERENDIP, a novel framework that implements this theory.

5. MODEL AND ALGORITHM

In this section, we elaborate the design of SERENDIP and present efficient inference algorithms to fit the model to real datasets using maximum likelihood estimation.

5.1 A Network Regularization Framework

For simplicity of presentation, we use the following notations: As usual, $\mathcal{P}_t$ denotes the set of papers published by the target organization at time t; $\mathcal{Q}_t$ now represents all the papers accessed by the target organization till t; $\mathcal{G}_t$ is the reference network constituting all the papers in the universe available by t. We intend to quantify the creativity observed in $\mathcal{P}_t$ in terms of both preparedness (explainable by $\mathcal{Q}_t$) and serendipity (unexplainable by $\mathcal{Q}_t$). In the following, when the context is clear, we omit the time subscript t.

To fulfill the model in Section 4, we need to decide: (i) the number of additional edges to be added to $\mathcal{G}$ and (ii) the set of nodes (i.e., papers) to be connected by these edges.

From Number of Edges to Age of Nodes

The number of edges, to a certain extent, captures the expected amount of preparedness to be explained by $\mathcal{Q}$. However, without prior knowledge, this quantity is fairly difficult to gauge. Thus, instead of directly specifying the number of edges expected to be added, we

6. APPLICATION

In this section we show that equipped with the aforementioned SERENDIP model, we are able to answer a set of fundamental yet challenging questions, including:

Prediction: Based on the materials consumed by the creators, can we predict what products they are going to

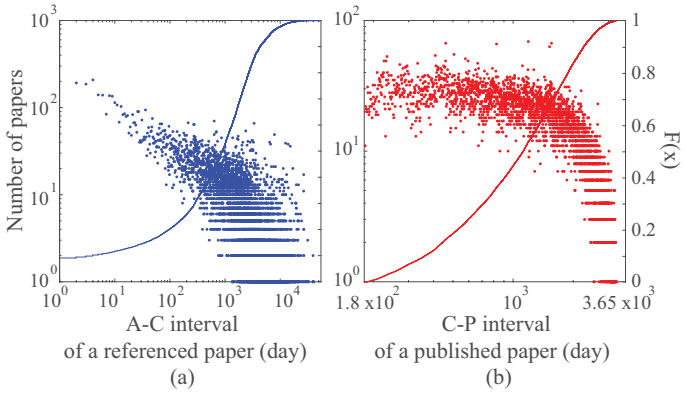

Figure 6: (a) Interval between when a referenced paper is published and when it is first accessed; (b) Interval between when one of its referenced papers is first accessed and when a paper is published.

produce? In terms of scientific research, what innovations are mostly like to happen by connecting the currently disparate knowledge?

Recommendation: Given the creators' target products, what are the most critical materials for them to consume? In terms of scientific research, which references are "must-read" for the authors to make their target innovations happen?

What-If Analysis: If the creators were exposed to certain materials, would they be able to produce much better products? In terms of scientific research, which references might help the authors significantly improve their current publications?

Next we detail how to answer these questions.

7. EMPIRICAL EVALUATION

7.1 A-C-P Life Cycle

Denote by $\mathcal{R}_{p^*}$ all the papers referenced by a given paper p^*. For each paper $p \in \mathcal{R}_{p^*}$, we consider three critical time points in its life cycle:

- *Availability* time $\mathsf{T}_p^{\mathsf{A}}$ - the time when p is published, thereby becoming available.
- *Consumption* time $\mathsf{T}_p^{\mathsf{C}}$ - the time when p is accessed (or consumed) by the target organization.
- *Publication* time $\mathsf{T}_p^{\mathsf{P}}$ - the time when p^* is eventually published, i.e., p is incarnated in p^*.

We then measure the following two temporal intervals:

- Availability-consumption interval $\Delta_p^{\mathtt{AC}} = \mathsf{T}_p^{\mathsf{C}} - \mathsf{T}_p^{\mathsf{A}}$, capturing the gap between when a referenced paper p becomes available and when it is actually consumed.
- Consumption-production interval $\Delta_p^{\mathtt{CP}} = \mathsf{T}_p^{\mathsf{P}} - \mathsf{T}_p^{\mathsf{A}}$, capturing the gap between when p is accessed and when it is incarnated in a published paper p^*.

Figure 6 shows the distributions of AC intervals and CP intervals in our dataset.

8. CONCLUSION

In this paper, we conducted an extensive empirical study on predicability in scientific creative processes. For the first time, by directly correlating authors' raw information consumption with their publications, we found remarkable reproductive patterns in scientific creativity across [#] authors. Furthermore, we proposed SERENDIP, a mechanistic modeling framework for scientific creative processes, which explicitly accounts for authors' information consumption. By using two Web-scale, longitudinal real datasets, we demonstrated the efficacy of SERENDIP in predicting disparate references most likely to be connected by creativity as well as identifying critical references necessary for such linkings to happen. SERENDIP is not limited to scientific creative processes. Indeed, the mechanistic nature of SERENDIP makes it potentially applicable for modeling creative processes in other domains of computational creativity, 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 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 impacts of scientific artifacts in their early stages. Hence, incorporating such semantic features in the creativity model could be integrated into SERENDIP model to train microscopic (author-, content-, and venue-specific) creativity models. Lastly, the SERENDIP 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.

9. 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] 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.

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

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

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

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

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

[8] 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.

[9] Y. Dong, J. Zhang, J. Tang, N. V. Chawla, and B. Wang. Coupledlp: Link prediction in coupled networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 2015.

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

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

[12] 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.

[13] 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.

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

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

[16] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, 2008.

[17] 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.

[18] L. Lu and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150 – 1170, 2011.

[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] J. Ni, H. Tong, W. Fan, and X. Zhang. Inside the atoms: Ranking on a network of networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, 2014.

[21] 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.

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

[23] 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.

[24] H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, 2006.

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

[26] 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.

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

[28] T. Wang, M. Srivatsa, D. Agrawal, and L. Liu. Microscopic Social Influence. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, SDM '12, 2012.

[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.

APPENDIX