Go Wide, Go Deep: Quantifying the Impact of Scientific Papers through Influence Dispersion Trees

Dattatreya Mohapatra[1*], Abhishek Maiti[1*], Sumit Bhatia[2] and Tanmoy Chakraborty[1]
[1]IIIT-Delhi, India; [2]IBM Research AI, New Delhi, India
{dattatreya15021,abhishek16005,tanmoy}@iiitd.ac.in,sumitbhatia@in.ibm.com

ABSTRACT

Despite a long history of the use of 'citation count' as a measure of scientific impact, the evolution of the follow-up work inspired by the paper and their interactions through citation links have rarely been explored to quantify how the paper enriches the depth and breadth of a research field. We propose a novel data structure, called Influence Dispersion Tree (IDT), to model the organization of follow-up papers and their dependencies through citations. We also propose the notion of an ideal IDT for every paper and show that an ideal (highly influential) paper should increase the knowledge of a field vertically and horizontally. We study the structural properties of IDT (both theoretically and empirically) and propose two metrics, namely Influence Dispersion Index (IDI) and Normalized Influence Divergence (NID) to quantify the influence of a paper. Our theoretical analysis shows that an ideal IDT configuration should have equal depth and breadth (and thus minimize the NID value). We establish the superiority of NID as a better influence measure in two experimental settings. First, on a large real-world bibliographic dataset, we show that NID outperforms raw citation count as an early predictor of the number of new citations a paper will receive within a certain period after publication. Second, we show that NID is superior to the raw citation count at identifying the papers recognized as highly influential through 'Test of Time Award' among all their contemporary papers (published in the same venue).

1 INTRODUCTION

A common consensus among the Scientometrics community is that the total number of citations received by a scientific article can be used to quantify its impact on the research field [16, 17]. Citation count, being a simple metric to compute and interpret, is commonly used in many decision-making processes such as faculty recruitment, fund disbursement, and tenure decisions. Many improvements over raw citation count have also been proposed by incorporating additional constraints. Examples include normalizing citation counts by the maximum citation count a paper could achieve in a particular research field [33], metrics inspired by PageRank [12], taking into account the locations of citation mentions in the paper (e.g. Introduction, Related Work, etc.) [37], understanding the reasons behind citations and assigning different weights to different citations based on these reasons [7].

While improvements over the raw citation count, these measures are fundamentally also *aggregate* measures as they ignore the relationships between different (citing) papers that cite a given paper. We posit that such connections are useful and studying them can help us better understand the propagation of influence from a paper to its different citing papers. Rather than proposing yet another variant of citation count, we are interested in unraveling these structural connections between the set of followup papers of a given paper and understand the differentiating structural properties of influential papers.

Motivation: We posit that the impact of a scientific paper can broadly be studied across two dimensions – *(i)* how many different research directions it gives rise to; and *(ii)* how much traction these individual research directions gather in the field. In the former case, we say that the influence of the paper has *breadth* and it helps in expanding the field horizontally, leading to an increase in the breadth of the field. A paper with such a broad influence may even trigger the emergence of a new sub-field. In the latter case, we say that the paper has had a *deep* influence on the field with a large number of papers in a given research direction. Intuitively, *highly influential papers are the ones that have a deep, and broad influence on the field.* Influence measures that are variants of the raw citation count of the paper may not offer such fine-grained understanding of the contribution of a paper to its field. Quantifying the impact of a paper in terms of its depth and breadth may also help to uncover the relationship between its different citing papers [24] and thus, understand the diffusion patterns of scientific ideas through citation links [9], predict the structural virality [19] and citation cascade [8, 24, 30]. While there have been recent efforts to study these structural properties of networks formed by a paper and its citing papers [24, 30], none of these studies have attempted to develop a metric to quantify the influence of a paper from its network topology. *We are the first to propose a series of metrics to quantify a new facet of influence that a paper has had on its followup papers.*

Our Contributions: Our major contributions are as follows.

(i) A framework to model the depth and breadth of the influence of a paper by a novel network structure, called the *Influence Dispersion Tree (IDT)* (Section 3). The IDT of a paper P is a directed tree rooted at P with all its citing papers as the children. The tree is constructed such that the citing papers having citation links among themselves are grouped to represent a body of work influenced by the root paper P (Section 3.1). These *bodies* of work along with the number of papers in each group are then used to model the depth and breadth of impact of P. We also present a theoretical analysis of the properties of the IDT structure and show how these properties are related to the citation count of the paper (Section 3.2).

(ii) A series of measures to quantify the influence of a scientific paper: For a scholarly paper P, we propose a novel metric, called *Influence Dispersion Index* (IDI) derived from its IDT to quantify the contribution of the paper to its field (by increasing depth or breadth or both) through influence diffusion (Section 3.3). We argue that in an ideal scenario, the influence of a paper should be dispersed to maximize the depth as well as the breadth of its influence. We then derive the configuration of the IDT of such a paper

*Equal contribution.

978-1-7281-1547-4/19/$31.00 ©2019 IEEE
DOI 10.1109/JCDL.2019.00051

and prove that such an optimal IDT configuration will have equal depth and breadth (and is equal to $\lceil \sqrt{n} \rceil$, where n is the number of citations of a given paper). Next, we propose another metric, called *Influence Divergence* (ID) that measures how the IDI value of a paper diverges from IDI value of the optimal IDT configuration (Section 3.5). A lower value of divergence indicates that the influence of the paper under consideration is dispersed in a way that is similar to that of the ideal case, and consequently, higher is the chance for the paper to be considered as a highly influential paper. We further derive a normalized version of ID, and call it *Normalized Information Divergence* (NID) that normalizes influence divergence values for different papers with different citation counts in the range $[0, 1]$ and allows for comparing different papers based on their NID values.

(iii) Empirical validation on large real-world datasets: We use a large bibliographic dataset consisting of about 3.9 million articles (Section 4) to study the properties of the proposed IDT structure and test the effectiveness of proposed influence metrics. We construct IDTs for all the papers in the dataset and their analysis reveals several interesting observations (Section 5). First, we observe that with an increase in the citation count, breadth of an IDT tends to grow much faster than the depth. The maximum value of breadth $(4,892)$ is much higher than that of depth (48). We infer that acquiring more citations over time often leads to an increase in the breadth instead of growth of an existing branch. Next, we find that the NID value decreases with an increase in citation count. This finding strengthens our hypothesis that the IDT of an highly influential paper tends to reach its optimal configuration by enhancing both the depth and the breadth of its research field. Third, we show that NID outperforms raw citation count as an early predictor to forecast the number of future citations a paper will receive (Section 6.1). Finally, we manually curate a set of 40 papers recognized as the most influential papers by their communities through 'Test of Time' or '10 years influential paper' awards. Once again, we find that NID outperforms the raw citation count in identifying these influential papers (Section 6.2). Most importantly, NID also provides an explanation why a paper has received such a prestigious award – it is not only the number of followup papers (or citation count) that matters, but the factor which affects most is the way the followup papers are organized and linked in an IDT. In other words, *a highly influential paper tends to have an IDT with high breadth as well as high depth*. For reproducibility, the code and the dataset are available at https://github.com/LCS2-IIITD/influence-dispersion.

2 RELATED WORK

There has been a plethora of research to measure the impact of scientific articles through various forms of citation analysis. In this section, we separate the related work into two parts – (i) studies dealing with citation count and its variants for measuring the impact, and (ii) studies exploring detailed orchestration of citations around scientific papers.

2.1 Citation Count as Impact Measure

Searching for accurate and reliable indicators of research performance has a long and often controversial history. Citation data is frequently used to measure scientific impact [16, 17]. Most citation indicators are based on citation counts – Journal Impact Factor [18], h-index [21], Eigenfactor [14], i-10 index [11], c-index [31], etc. Many variations and adaptations were proposed to compensate the drawbacks of these indices. For instance, m-quotient [21, 39] attempts to eliminate the bias of h-index towards older researchers/articles. g-index [13] and e-index [41] were proposed to overcome bias again authors with heavily cited articles. We proposed C^3-index [32] to resolve ties while ranking medium-cited and low-cited authors by h-index. Even though so many variations of h-index were proposed in the literature, Bornmann et al. [4] concluded that most of them are redundant by showing a mean correlation coefficient of 0.8-0.9 between h-index and its 37 alternatives. Few attempts were made to quantify the contribution of individual authors in multi-authored publications [23, 25, 27, 36].

To measure the impact of a scientific article, raw citation count has by far been the most accepted and well studied metric [33, 35]. However, many studies confronted with different views against citation count, giving rise to several alternatives such as *influmetrics* [3], *webometrics* [1], *usage metrics* [26], *altmetrics* [20], etc. Chakraborty et al. [5] showed that the change in yearly citation count of articles published in journals is different from articles published in conferences. Even the evolution of yearly citation count of papers varies across disciplines [6, 34]. This further raises a new proposition of designing domain-specific impact measurement metrics.

2.2 Understanding Citation Expansion

Despite such a vast literature on the use of citation count for assessing the quality of scientific community, the evolution of citation structure has remained largely unexplored. There have been a few recent studies which attempted to understand the organization of citations around a scientific entity (paper, author, venue etc.), particularly focusing on the topology of the graph constructed from the induced subgraph of papers citing the seed paper. Waumans and Bersini [40] took an evolutionary perspective to propose an algorithm for constructing genealogical trees of scientific papers on the basis of their citation count evolution over time. This is useful to trace the evolution of certain concepts proposed in the seed paper. Singh et al. [38] developed a relay-linking model for prominence and obsolescence to include the factors like aging, decline etc. in the evolving citation network. Min et al. [29] characterized the citation diffusion process using a classic marketing model [2] and noticed some interesting patterns in the spread of scientific ideas. Inspired by information cascade modeling in online social networks [10], they [30] further made an attempt to study the behavior of citation cascade. They concluded that the average depth of the cascade tends to be influenced by both the lifespan and the whole volume of scientific literature. Huang et al. [24] and Chen [8] argued that citation cascade helps us better understand the citation impact of a scientific publication. They empirically showed that most of the properties of the cascade graph (such as cascade size, edge count, in-degree, and out-degree) follow typical power law distributions; however cascade depth follows exponential distribution.

2.3 Differences from Previous Literature

Although recent studies [8, 24, 30] argued that there is a need to explore the organization of citations (followup papers) around a seed paper in order to measure better scientific impact, no one

quantitatively studied the impact of such network. We are the first to propose an impact measurement metric, called 'Influence Dispersion Index' (Section 3.3) which is derived upon converting a rooted citation network to a sparse representation, called 'influence dispersion tree' (IDT) (Section 3). We show how an optimal orientation of CDT (in terms of its depth and breadth) helps in gaining more impact, which may not be explained by simple citation count. Moreover, the construction of IDT is unique and different from the citation cascade graph proposed earlier [8, 24, 30] (see Section 3 for more details).

3 INFLUENCE DISPERSION TREE (IDT)

In this section, we first develop and define the concept of Influence Dispersion Tree of a scholarly paper and describe some of the properties of IDTs. We then develop a simple measure to estimate the *influence* of a scholarly paper given its IDT.

3.1 Constructing IDT

Let us consider a scholarly paper P and let $C_P = \{p_1, p_2, \ldots, p_n\}$ be the set of papers citing P. We assume that P has *equally and directly* influenced each and every paper in C_P.[1]

Definition 1. [**Influence Dispersion Graph**] The Influence Dispersion Graph (IDG) of the paper P is a directed and rooted graph $\mathcal{G}_P(\mathcal{V}_P, \mathcal{E}_P)$ with $\mathcal{V}_P = C_P \cup \{P\}$ as the vertex set and P as the root. The edge set $\mathcal{E}_P$ consists of edges of the form $\{p_u \rightarrow p_v\}$ such that $p_u \in \mathcal{V}_P, p_v \in C_P$ and p_v cites p_u.

Figure 1(a) shows an illustration of an IDG for the paper P and its citing paper set $\{p_1, p_2, p_3, p_4, p_5\}$. Observe that the IDG of paper P is the same as the induced subgraph of the larger citation graph consisting of P and all its citing papers, and with edges in the opposite direction to indicate the propagation of influence from the cited paper to the citing paper. Further, note that the construction of an IDG is similar to that of citation cascades [24] with the fundamental difference that the IDG is restricted strictly to the one-hop citation neighborhood of P (i.e., papers that are directly influenced by P) as opposed to the citation cascade that considers higher order citation neighborhoods as well (i.e., papers indirectly influenced by P). Thus, an IDG only considers followup papers that are *directly influenced* by a given paper. If p_1 cites P; and p_2 cites p_1 but not P, it is not always clear if p_2 is influenced by both P and p_1, or solely by p_1. Thus, we make the stricter and unambiguous choice by selecting only p_1 to be included in the IDG. Though variants of IDG could be constructed by adding additional followup papers, we believe that the major conclusions drawn from the paper will remain valid owing to the stricter and unambiguous process of constructing the IDG.

Next, to further analyze and study the influence of paper P on its citing papers, we derive the *Influence Dispersion Tree* (IDT) of P from its IDG. A tree structure, by definition, provides a hierarchical view of the influence P exerts on its citing papers and provides an easy to understand representation to study the relation between P and its citing papers. The IDT of paper P is a directed and rooted tree $\mathcal{T}_P = \{\mathcal{V}_P, \mathcal{E'}_P\}$ with P as the root. The vertex set is the same as that of IDG of P and the edge set $\mathcal{E'}_P \subset \mathcal{E}_P$ is derived from the edge set of IDG as described next.

Note that a paper $p_v \in C_P$ can cite more than one paper in $\mathcal{V}_P$, giving rise to the following three possibilities:

(1) p_v cites only the root paper P. In this case, we add the edge $P \rightarrow p_v$ creating a new branch in the tree emanating from root node (e.g., edges $P \rightarrow p_1$ and $P \rightarrow p_2$ in Fig. 1(b)).
(2) p_v cites the root paper P and $p_u \in C_P \setminus \{p_v\}$. In this case, we say that p_v is influenced by P as well as p_u. There are two possible edges here: $P \rightarrow p_v$ and $p_u \rightarrow p_v$. However, since p_u is also influenced by P, the edge $p_u \rightarrow p_v$ indirectly captures this influence that P has on p_v. We therefore retain only the edge $p_u \rightarrow p_v$. This choice leads to addition of a new leaf node in IDT capturing the chain of impact starting from P up to the leaf node p_v (e.g., edge $p_1 \rightarrow p_3$ in Fig. 1(b)).
(3) p_v cites the root paper P, as well as a set of other papers $P_u \subseteq C_P \setminus \{p_v\}$, $|P_u| >= 2$. Note that by definition, each $p \in P_u$ also cites the root paper P. The possible edges to add here are $E = \{\{p \rightarrow p_v\}; \forall p \in P_u\}$. We add the edge e to $\mathcal{E}'_P$ such that $e = p \rightarrow p_v$ where

$$p = \arg\max_{p' \in P_u} shortestPathLength(P, p') \quad (1)$$

Edge $P_3 \rightarrow P_5$ in Fig. 1(b) is such an edge.

The intuition behind adding edges in this way is to maximize the depth of IDT (if there are more than one edge, and each of which maximizes the depth, then we choose one of them randomly, e.g., $p_2 \rightarrow p_4$ in Fig. 1(b)). The edge construction mechanism is motivated by the citation cascade graph [24, 30]. Upon adding a newly citing paper in $\mathcal{T}_P$, we reconstruct $\mathcal{T}_P$ in such a way that the richness of P's influence to its citing papers is maximally preserved. Richness maximization can be thought of as maximizing the breadth or the depth of the IDT. We choose the latter one in order to capture the cascading effect into the resultant IDT.

Definition 2 (**Influence Dispersion Tree).** The Influence Dispersion Tree (IDT) of paper P is a tree $\mathcal{T}_P(\mathcal{V}_P, \mathcal{E'}_P)$, whose vertex set $\mathcal{V}_P$ is the union of P and all the papers citing P. If a paper p_v cites only P and no other papers in $\mathcal{V}_P$, we add $P \rightarrow p_v$ into the edge set $\mathcal{E}'_P$. If p_v cites other papers $P_u \in \mathcal{V}_P \setminus \{P\}$ along with P, we add only one edge $p_x \rightarrow p_v$ (where $p_x \in P_u$) according to Equation 1.

Definition 3 (*P*-**rooted IDT).** An IDT is called P-rooted IDT when the root node of the tree is P.

Figure 1 illustrates a toy example of constructing IDT from IDG illustrating all three possible cases of edge connections as discussed above.

3.2 Properties of IDT

In this section, we describe a few important properties of an IDT.

(i) Depth: The depth d of a P-rooted IDT is defined as the length of the longest path from the root to the leaf nodes p_L in the tree.

$$d = \max_{p_l \in p_L} shortestPathLength(P, p_l) \quad (2)$$

[1] Although previous studies [7, 42] have found that a paper has a varying amount of influence on its citing papers, it is a common practice to assume uniform influence for simplification (e.g., in computing impact factors, h-index [22], etc.) and is the assumption we also make.

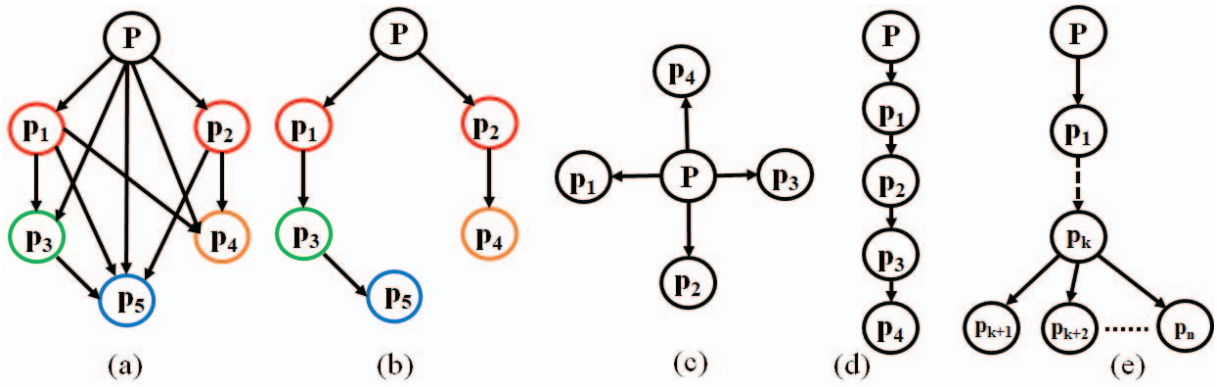


Figure 1: (a)-(b) Illustration of the construction of (b) IDT from (a) IDG of paper P. Papers in red only cite P; Papers in green cite P and one other paper in the graph; blue paper cites P and more than one other paper in the graph. In case of yellow paper, a tie-breaking occurs due the equal possibility of p_4 being connected from p_1 and p_2 in order to maximize the depth of IDT. Tie-breaking is resolved by randomly connecting p_4 from p_2 in IDT. (c)-(d) Two corner cases to illustrate the lower bound – minimum and maximum number of leaf nodes. (e) A configuration of a P-rooted IDT with (n) non-root nodes that results in maximum IDI value.

where d is the depth of the tree, and p_L is the set of leaf nodes in IDT. The depth of the IDT shown in Figure 1(b) is 3.

The depth of an IDT can be interpreted as the longest chain/series of papers representing a body of work influenced by P.

(ii) Breadth: The breadth b of a P-rooted IDT is defined as the maximum number of nodes at a given level in the tree.

$$b = \max_{1 \le l \le d} |N_l|; \quad N_l := \{n \in \mathcal{V}_P | level(n) = l\} \tag{3}$$

The breadth of the IDT shown in Figure 1(b) is 2.

(iii) Branch: A branch $P \rightsquigarrow p_l$ is a path from the root P to the leaf p_l in an IDT.

(iv) Fragmented and Unified Branch: A branch $P \rightsquigarrow p_l$ is called fragmented when an intermediate node (except root) $p \in P \rightsquigarrow p_l$ becomes a part of another branch $P \rightsquigarrow p_{l'}$. p is then called a **fragment point** of $P \rightsquigarrow p_l$. In Figure 1(e), $P \rightsquigarrow p_{k+1}$ is a fragmented branch with p_k as a fragment point. If a branch is not fragmented, it is called as a unified branch. In Figure 1(d), $P \rightsquigarrow p_4$ is a unified branch.

We now define some properties to describe how depth and breadth of a P-rooted IDT are related with n – the number of citations of P (and the number of non-root nodes in the IDT of P).

Lemma 1. *For a paper P with n citations, the range of the depth d and breadth b of the P-rooted IDT is $1 \le d, b \le n$.*

Proof. The breadth of a P-rooted IDT will be maximum (i.e, n) when all the n papers cite only the root paper P, and there is no citation among these n papers (e.g. Figure 1(c)). Likewise, the depth of a P-rooted IDT will be maximum (i.e., n) when there is a chain of n papers $\{P, p_1, p_2, \cdots, p_n\}$ forming a unified branch such that p_i cites p_{i-1}, $\forall 2 \le i \le n$; and p_i also cites P, $\forall i$ (e.g., Figure 1(d)). □

Lemma 2. *For a paper P with n citations, the sum of depth d and breadth b of the P-rooted IDT is bounded by $n+1$, i.e., $d + b \le n + 1$.*

Proof. When a new node is added to IDT, there are four possibilities – breadth increases, depth increases, both increase, and neither increases. The sum of d and b will be maximum when both of them are individually maximum. This will only be possible when all but the root node are involved in either increasing depth or breadth or both. However, we can see that only one node, i.e., the first node attached to the root node, can increase both depth and breadth, and the rest will increase either depth or breadth, but not both. Since the total number of non-root nodes added to IDT are n, the sum of b and d can attain a maximum value of $n + 1$. □

Lemma 3. *For a paper P with n citations and its P-rooted IDT, the product of its depth d and breadth b is at least n, i.e., $db \ge n$*

Proof. d is the maximum length of any branch, and b is indicative of the number of branches from root to leaf. So, for an IDT whose branching occurs at the root node itself and nowhere else, db represents the number of nodes it can have to maintain its depth as d and breadth as b by adding to those branches which have less than d length. Since n is the number of nodes already present in the IDT, we can say that the number of nodes we can add is $db - n$. Since this quantity is always non-negative as this quantity represents the number of nodes we can add, we have

$$db - n \ge 0 \implies db \ge n \tag{4}$$

For those IDTs which have branching in places other than the root i.e., fragmented branches, the nodes which are above the branching nodes, will be counted more than once as they represent multiple root to leaf paths and hence db will give more number of nodes than present in the IDT; hence

$$db > n \tag{5}$$

Therefore, for both the cases, it is seen that $db \ge n$. □

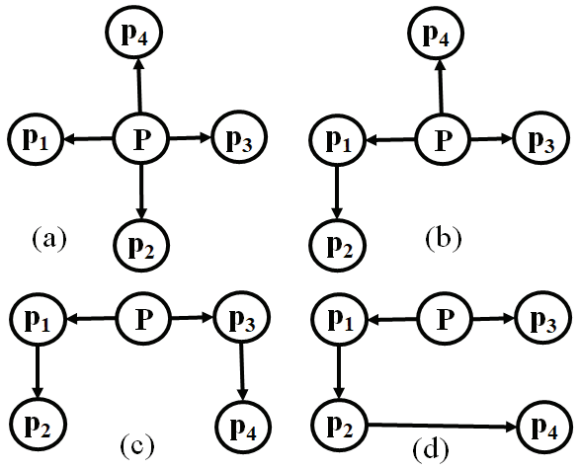


Figure 2: Reconnecting leaf edges of a star IDT (a) to form other configurations.

3.3 Influence Dispersion Index (IDI)

Given the IDT of a paper, we define its Influence Dispersion Index (IDI) by the sum of length of all the paths from the root node to all the leaf nodes.

Definition 4 (Influence Dispersion Index). The IDI of paper P is defined as

$$IDI(P) = \sum_{p_l \in p_L} distance(P, p_l) \tag{6}$$

where p_L is the set of leaf nodes of the P's IDT $\mathcal{T}_P(\mathcal{V}_P, \mathcal{E}_P)$.

The IDI of P in Figure 1(b) is 5.

Intuitively, each leaf node in P's IDT corresponds to a separate branch emanating from the original paper P. Each branch comprises of the set of papers which are influenced by the root paper in one direction. We can interpret IDI as a measure of the *ability* of the paper to distribute its influence. We hypothesize that the more an IDT has unified branch, the more the chance that the influence emanating from P is distributed uniformly.

3.4 Boundary Conditions of IDI

3.4.1 Lower Bound. For a P-rooted IDT with n non-root nodes, the minimum value of IDI is n. This is because each node (paper) in the tree will be encountered at least once while computing IDI, resulting in the lower bound as n. Figures 1(c) and (d) show two corner cases – one configuration with the minimum number of leaf nodes (i.e, 1), and other configuration with the maximum number of leaf nodes (i.e., n). Note that given the size of the IDT, there can be multiple configurations with minimum IDT values. From a star IDT (Figures 1 (c)) if we pick an edge and connect it to any leaf node or the root node, then IDI of the resultant configuration will remain same. In fact, if we keep on repeating the same repairing step, all the resultant configurations will exhibit the same IDI value. In short, during the transformation of a star IDT to a line IDT by reconnecting a leaf edge (an edge whose one end node is a leaf) to another leaf node or to the root node, all the intermediate IDTs will exhibit the same IDI of n. Figure 2 shows a toy example of the reconfiguration. We will discuss more in Section 3.4.3.

3.4.2 Upper Bound: In order to maximize the value of IDI, a P-rooted IDT should satisfy the following three conditions:

(1) The number of leaves should be as large as possible.
(2) The length of the branch from root to leaf should be as long as possible.
(3) The number of common nodes in each root-to-leaf branch should be maximized so that each node counter is maximized.

Subject to the constraint on the number of nodes in the tree (i.e., $n+1$), there is only one structure which can satisfy all the three requirements mentioned above, as shown in Figure 1(e).

Let IDI of the P-rooted IDT with n non-root nodes as shown in Figure 1(e) be $IDI(P,k)$, where k is the number of nodes forming a chain from P (excluding P) and node p_k has $(n-k)$ descendants. Then, $IDI(P,k)$ is determined as follows:

$$IDI(P,k) = k(n-k) + (n-k) \tag{7}$$

Differentiating it w.r.t to k, we get

$$\frac{\partial IDI(P,K)}{\partial k} = n - 2k - 1 \tag{8}$$

Equating this to 0 to get the maxima, we get

$$k = \left\lfloor \frac{n-1}{2} \right\rfloor \tag{9}$$

This yields the maximum value of IDI as

$$IDI(P)^{max} = (1 + \left\lfloor \frac{n-1}{2} \right\rfloor)(n - \left\lfloor \frac{n-1}{2} \right\rfloor) \tag{10}$$

Therefore, for a P-rooted IDT with n non-root nodes, we have the following bounds on its IDI:

$$\boxed{n \quad \leq \quad IDI(P) \quad \leq \quad (1 + \left\lfloor \frac{n-1}{2} \right\rfloor)(n - \left\lfloor \frac{n-1}{2} \right\rfloor)} \tag{11}$$

3.4.3 Relation between d, b and n for Optimal Dispersion. As discussed above, a paper with a given number of citations n, can have differently shaped IDTs, and consequently, very different IDI values. Intuitively, we expect a highly influential paper to have multiple long unified branches, i.e., *it should have a high depth value as well as high breadth value.* Thus, we want the IDT of a highly influential paper to have high depth, high breadth, and a tree structure such that the number of non-root nodes are as uniformly distributed in different branches of the trees as possible, indicating significant depth in each branch. Also, recall from Lemma 3 that for a given value of d and b, the number of nodes in an IDT can not be more than db (i.e., $n \leq db$). This leads us to the following constrained objective function that the IDT in its optimal configuration should satisfy.

$$\begin{aligned} &minimize && (db - n) && \\ &\text{s.t} && d + b \leq n + 1 && \text{(from Lemma 2)} \\ &\text{and} && db \geq n && \text{(from Lemma 3)} \end{aligned} \tag{12}$$

This yields an optimal configuration where $d = b = \lfloor \sqrt{n} \rfloor$.

Proof. As discussed, db represents the maximum number of nodes the tree can have by having depth as d and breadth as b. The IDT will have maximum number of nodes for a given d and b

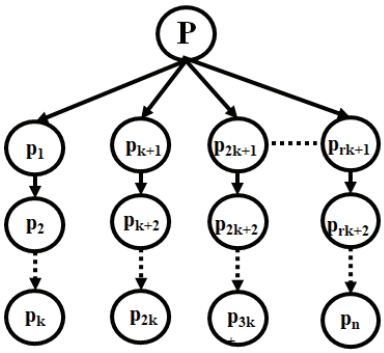


Figure 3: Illustration of an optimal configuration of a P-rooted IDT of a paper P with n citations. The depth and breadth of the IDT are same ($k = r = \lceil\sqrt{n}\rceil$).

only when all the branches in the IDT are unified branches. This condition will force the IDT to have all the branches to branch out from the root node. If k is the number of nodes in each unified branch of the optimal tree, and there are r such branches, then the number of nodes in this IDT will be kr (assuming equal length for each branch). Since k and r are equal for an optimal IDT as discussed earlier, we have

$$k^2 = n \Rightarrow k = \sqrt{n} \tag{13}$$

For IDTs where the nodes are not evenly distributed among an equal number of unified branches with each branch having equal number of nodes (in other words, when the number of non-root nodes is not a perfect square), the corresponding k comes out to be

$$k^2 = n \Rightarrow k = \lceil\sqrt{n}\rceil \tag{14}$$

□

Figure 3 illustrates a paper with an optimal configuration where the IDT has an equitable distribution in terms of both depth and breadth, indicating that the paper has influenced multiple branches, and all the influenced branches have grown significantly. Note that the cost function favors configurations where the impact of the paper is maximized both in terms of depth and breadth, and hence, will penalize configurations where there exists a large number of short branches (high b, low d) or very few long branches (high d, low b).

3.5 IDI as an Influence Measure

In this section, we study the potential of IDI as an early predictor of the overall impact and influence of a scholarly article. As discussed before, IDI of a paper P provides a fine-grained view of the influence of P on other papers citing P, in terms of the depth and breadth of the IDT. As described in Section 3.4, for a paper with n citations, there exists an ideal configuration of the IDT that optimizes the influence dispersion of the paper such that it has both high breadth (influenced multiple branches of work) and high depth (significantly deepened each individual branch). With this intuition, we posit that the *closeness* of the actual IDT of a given paper P with n citations, denoted by $\mathcal{T}_P$ to its corresponding ideal IDI with n citations, denoted by $\bar{\mathcal{T}}_P$ can be used as a surrogate measure of influence or impact of paper P. We can use any distance metric between two graphs – such as Graph Edit Distance [15], Gromov-Wasserstein distance [28] – to measure the closeness between $\mathcal{T}_P$ and $\bar{\mathcal{T}}_P$. However, all these measures are computationally expensive [15]. Therefore, we here use the IDI of each IDT as a proxy for its topological structure and measure the difference between the IDI values of $\mathcal{T}_P$ and $\bar{\mathcal{T}}_P$ (as a replacement of the graph distance). Recall from Section 3.4 that the IDI of an ideal IDT with n non-root nodes is n (which is also the lower bound of an IDT with n internal nodes).

We define the **Influence Divergence** (ID) of a paper as the difference of the IDI value of its original IDT, IDI(P) and that of its corresponding ideal IDT configuration, $I\bar{D}I$(P)

$$ID(P) = IDI(P) - I\bar{D}I(P) \tag{15}$$

We further normalize the IDI value using max-min normalization.

Definition 5 (Normalized Influence Divergence). Normalized Influence Divergence (NID) of a paper P is defined by the difference between the IDI value of its corresponding IDT and the same of its corresponding ideal IDT configuration, $I\bar{D}I$(P), normalized by the difference between maximum and minimum IDI values of the IDTs with the size as that of P's IDT. Formally, it is written as:

$$NID(P) = \frac{IDI(P) - I\bar{D}I(P)}{IDI^{max}_{|P|} - IDI^{min}_{|P|}} \tag{16}$$

The normalization is needed to compare two papers with different IDI values. NID ranges between 0 and 1. Clearly, a highly influential paper will have a low $NID(P)$ (i.e., lower deviation from its ideal dispersion index).

4 DATASET DESCRIPTION

We used a publicly available dataset of scholarly articles provided by Chakraborty and Nandi [6]. The dataset contains about 4 million articles indexed by Microsoft Academic Search (MAS)[2]. For each paper in the dataset, additional metadata such as the title of the paper, its authors and their affiliations, year and venue of publication are also available. The publication years of papers present in the dataset span over half a century allowing us to investigate diverse types of papers in terms of their IDTs. A unique ID is also assigned to each author and publication venue upon resolving the named-entity disambiguation by MAS itself. We passed the dataset through a series of pre-processing stages such as removing papers that do not have any citation and reference, removing papers that have forward citations (i.e., citing a paper that is published after the citing paper; this may happen due to archiving the paper before publishing it), etc. This filtering resulted in a final set of 3, 908, 805 papers. Table 1 shows different statistics of the filtered dataset.

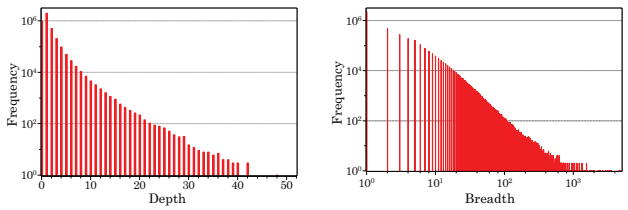
5 EMPIRICAL OBSERVATIONS

In this section, we report various empirical observations about the IDTs of the papers in our dataset that provide a holistic view of the topological structure of the trees. We also study the how depth and breadth of the IDTs, the IDI and NID values vary with the citation count of the papers.

[2]https://academic.microsoft.com/

Number of papers	3,908,805
Number of unique venues	5,149
Number of unique authors	1,186,412
Avg. number of papers per author	5.21
Avg. number of authors per paper	2.57
Min. (max.) number of references per paper	1 (2,432)
Min. (max.) number of citations per paper	1 (13,102)

Table 1: Some important statistics about the MAS dataset.



(a) Depth **(b) Breadth**

Figure 4: Frequency distributions for depth (4a) and breadth (4b) of IDTs of all the papers in the dataset. The x-axis in the plot for breadth is in logarithmic scale.

5.1 Structural Properties of IDTs

Figure 4 plots the frequency distribution of depth and breadth of the IDTs for all the papers in the dataset. Observe that the values for breadth follow a very long tail distribution with about 75% of papers having a breadth less than or equal to 3 (note the log-scale on x-axes in Fig. 4b). On the other hand, the range of the depth values for IDTs is much smaller compared to the range of breadth values. The maximum value of depth is 48 compared to the maximum breadth of 4, 892. To illustrate the types of papers that achieve very high breadh and depth values, Table 2 lists the top two papers having maximum depth (Papers 1 and 2) and maximum depth (Papers 3 and 4) in our dataset. Note that Papers 1 and 2 are famous Computer Science textbooks resulting in such high breadth values as most of the citing papers of a book (or survey papers) usually cite the book as a background reference. This may lead to a large number of short branches in the IDT. On the other hand, Papers 3 and 4 correspond to breakthrough seminal papers – Paper 3 was among the first to discuss and propose a solution for control flow problem in TCP/IP networks, and Paper 4 is Codd's seminal paper introducing relational databases. These groundbreaking works led to multiple followup papers that build upon these papers resulting in very high depth and relatively low breadth. Also note that even though Papers 3 and 4 have relatively fewer citations than Papers 1 and 2, analyzing the IDT enables us to *understand the depth and breadth of the impact of these papers on their citing papers* and measure the influence these papers have had on the fields.

Figure 5 shows the distribution of breadth and depth with citations (Figures 5a and 5b, respectively) and the correlation between depth and breadth (Figure 5c). We observe that while breadth is strongly correlated with citation count ($\rho = 0.90$), the correlation between depth and citation count is relatively weak ($\rho = 0.50$). These observations indicate that increasing citation count often lead to the development of new branches in the IDT of the paper rather than increasing the depth. This happens because most citations to a paper use the cited paper as a background reference (thus gets added to the IDT as a new branch), rather than extending a body of work represented by an already formed branch (increasing the depth). Further, note from Figure 5c that the variation in breadth values reduces with increasing depth. Especially for IDTs with depth greater than 30, the values of breadth lie in a relatively narrow band (almost all IDTs with depth greater than 30 have breadth less than 300). This is indicative of highly influential papers that have spawned multiple directions of follow-up works and incremental citations correspond to continuation of these independent directions (thus increasing depth).

5.2 IDI and NID vs. Citations

We now study how the IDI and NID values vary with the citation counts across multiple papers. Figure 6 shows the scatter plot of IDI and NID values with citations for all the papers in the dataset. We observe that IDI values in general increase with the number of citations of a paper. This is along expected lines as the IDI for a paper is bounded by the number of citations of the paper (Equation 11). A more interesting observation can be made from the plot for NID values (Figure 6b) where we see that in general, the value of NID decreases with increasing citations – papers having a high number of citations tend to have very low values of NID. Recall that for a given paper, NID captures how *different* or *far way* the IDI of the given paper is from its corresponding ideal IDT. Thus, highly influential papers tend to have their IDTs close to their ideal IDT configurations (as illustrated by the low NID value). This empirical observation strengthens our hypothesis that *highly influential papers will, in general, lead to considerable amount of followup work (high depth) in multiple directions (high breadth).*

6 NID AS AN INDICATOR OF INFLUENCE

As discussed before, we hypothesize that the highly influential papers produce IDTs which would be close to their corresponding ideal configurations. In Section 5.2, we found that highly-cited papers have very low NID values. Here we ask a complementary question – *Is low IDI value of a given paper an indicator of its future influence?* In other words, does a paper having its IDT close to the ideal configuration at a given time will be an influential paper in near future? We design two experiments to answer the above question. In Section 6.1, we study if NID can predict how many citations a paper will get in future. In Section 6.2, we study if IDI measure can identify highly influential papers – specifically, papers that have been judged highly influential by the community and have been awarded Test of Time (ToT) awards[3].

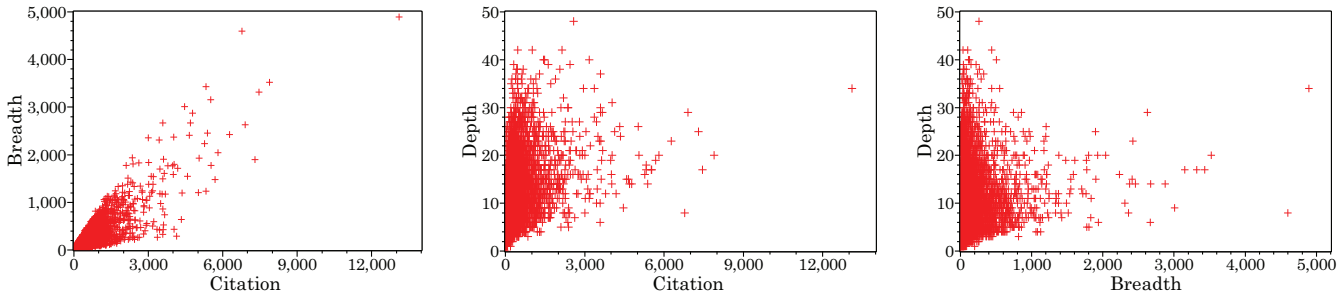
6.1 Future Citation Prediction through NID

Let P_v be the set of papers published in a publication venue v (a conference or a journal). Let y_v be the year of organization of v. Over the next t years, papers in P_v will influence the follow up

[3]Many conferences and journals award 'Test of Time' or '10 year influential paper award' to papers that have had a high impact on their respective fields. These papers are generally selected by a committee of senior researchers.

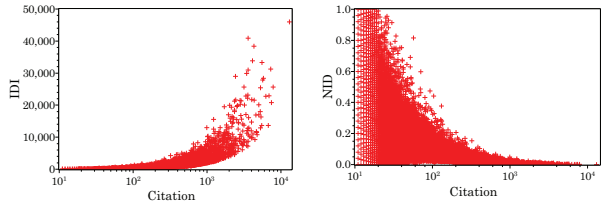
No.	Paper	# citations	breadth	depth	Remark
1.	Michael R. Garey and David S. Johnson. 1990. Computers and Intractability; a Guide to the Theory of NP-Completeness. W. H. Freeman & Co., New York, NY, USA.	13,102	4,892	34	A book on the theory of NP-Completeness
2.	Cormen, Thomas H., et al. (2001) Introduction to algorithms second edition.	6777	4576	8	Highly referred text book on Algorithms.
3.	CV. Jacobson. 1988. Congestion avoidance and control. In Symposium proceedings on Communications architectures and protocols (SIGCOMM '88), New York, NY, USA, 314-329.	2,577	259	48	Highly influential paper describing Jacobson's algorithm for control flow in TCP/IP networks
3.	E. F. Codd. 1970. A relational model of data for large shared data banks. Commun. ACM 13, 6 (June 1970), 377-387.	2141	437	42	Codd's Seminal paper on Relational Databases

Table 2: A set of representative papers: #1 and #2 are the top two papers based on breadth, and #3 and #4 are the top two papers based on depth.



(a) Breadth vs. Citations **(b) Depth vs. Citations** **(c) Depth vs. Breadth**

Figure 5: Scatter plots showing variations of breadth with citations (a), depth with citations (b), and correlation between depth and breadth (c).



(a) IDI vs. Citations **(b) NID vs. Citations**

Figure 6: Scatter plots showing variations of (a) IDI and (b) NID values with citation counts.

work and will gather citations accordingly. Let $I(p)$ be an influence measure under consideration. Let $R(v, t, I)$ be the ranked list of papers in P_v ordered by the value of $I(.)$ at t. Thus, the top ranked paper in $R(v, t, I)$ is considered to have maximum influence at t. If $I(.)$ is able to capture the impact correctly, we expect the papers with high influence scores to have more incremental citations in future compared to papers having low influence scores. Let $C(v, t_1, t_2)$ be the ranked list of papers in P_v ordered by the increase in citations from time t_1 to t_2. Thus, the papers that received highest fractional increase in citations in the time period (t_1, t_2) will be ranked at the top. Note that we chose fractional increase in citation count rather than absolute count to account for papers that are early risers and receive most of their lifetime citations in first few years after publication [5]. Also, we consider only those papers published in a venue (v here) rather than all the papers in our dataset to nullify the effect of diverse citation dynamics across fields and venues [6].

Intuitively, if $I(.)$ is a good predictor of a paper's influence, the ranked lists $R(v, t_1, I)$ and $C(v, t_1, t_2)$ should be very similar – influential papers at time t_1 should receive more incremental citations from t_1 to t_2. Thus, the similarity of the two ranked list could be used as a measure to evaluate the potential of $I(.)$ to be able to capture the influence of papers. We use the Kendall Tau rank distance $\mathcal{K}$ defined below to measure the similarity of the two ranked lists $R(v, t_1, I)$ and $C(v, t_1, t_2)$ as follows.

$$z(v, I) = \mathcal{K}(R(v, t_1, I), C(v, t_1, t_2)) \tag{17}$$

A lower value of the z score indicates that the two ranked lists are highly similar, that in turn shows that $I(.)$ has high predictive power in forecasting the future incremental citations. We use this framework to evaluate the potential of NID (as a replacement $I(.)$ in this case) as an early predictor of future incremental citations of a paper. We use the number of citations of a paper as a competitor of NID as it is the most common and simplest way of judging the influence of a paper [16, 17]. First, we group all the papers in our dataset by their venues and compute the values of the influence metrics (NID and citation count) after five years following the publication year (i.e., $t_1 = 5$). A venue is uniquely defined by the year of publication and the conference/journal series. For example, JCDL 2000 and JCDL 2001 are considered as two separate venues. We next compute the incremental citations gathered by the papers ten years after the publication ($t_2 = 10$). Note that we only consider

venues with the publication year in the range 1995 and 2000 because we needed citation information 10 years after publication (i.e., up to 2010). The coverage of papers published after year 2010 is relatively sparse in our dataset [6]. This filtering resulted in 1,219 unique venues and 30,556 papers in total.

With the group of papers published together in a venue and their citation information available, we compute the following three ranked lists:

(1) $R_{v,c} = R(v, 5, c)$; the ranked lists of papers in venue v ordered by their citation counts five years after the publication.
(2) $R_{v,nid} = R(v, 5, nid)$; the ranked lists of papers in venue v ordered by their NID scores five years after the publication.
(3) $C_v = C(v, 5, 10)$; the ranked lists of papers in venue v ordered by the normalized incremental citations received beginning of 5^{th} years after the publication till 10^{th} years after publication.

For each venue v, these lists can be used to compute $z(v, NID)$ and $z(v, c)$ – i.e., the z scores with NID and citation count as influence measures, respectively. For the 1,219 venues identified as above, the average value of z score using citations and IDI as the influence measure is found to be 0.5125 and 0.3703. Thus, on an average, we find that the Z score is lower when using NID as the influence measure compared to that with citation count. In other words, more papers identified as influential by NID received more incremental future citations compared to the papers identified as influential by citation count.

Figure 7 provides a fine-grained illustration of the difference of z scores achieved by the two influence measures for each of the 1,219 venues. For each venue, we compute the difference of z scores achieved by NID and citation count. We note that for most of the venues, the z-score achieved by NID is lower than the z-score achieved by the citation count (positive bars). These observations indicate that when compared with raw citation count, NID is a much stronger predictor of the future impact of a scientific paper. As opposed to the raw citation count, the IDT of a paper provides a fine-grained view of the impact of the paper in terms of its depth and breadth as succinctly captured by the IDT of the paper. These results provide compelling evidence for the utility of IDT (and the consequent measures such as IDI and NDI derived from it) for studying the impact of scholarly papers.

6.2 Identifying Test of Time Winners

Many conferences recognize highly influential papers that have had a long-lasting impact on the respective field of research. These recognitions are awarded in the form of Test of Time (ToT) awards, 10 year Influential Paper Awards, etc. We manually collected a set of papers that have received the ToT awards by their respective publication venues and obtained a list of 40 such papers (published in conferences like SIGIR, AAAI, ICCV etc.) that are also present in our dataset.

Let P be a ToT awardee paper that was published in year y at venue v. We extracted all the papers from our dataset that were published at venue v in year y. We then ordered these papers by their citation count at time $y + 10$ (i.e., 10 years after publication) and selected top 5% highest-cited papers (including P). We consider these papers to be the major competitor of P to win the TOT

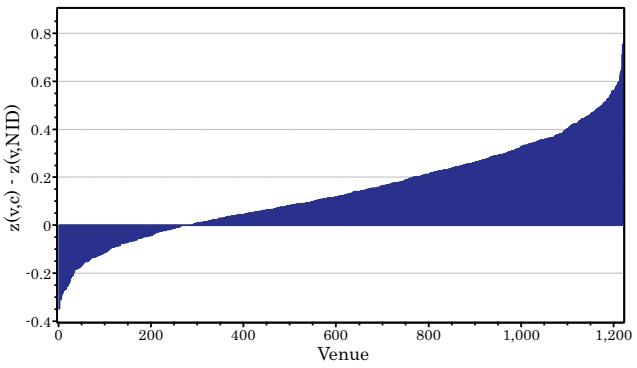


Figure 7: z-scores for venues. Papers in a venue are ranked using NID, number of citations and relative gain in citations. The horizontal axis represents venues ordered by the difference in two z-scores.

award since highly influential papers are expected to achieve a high number of citations[4]. We then compute the rank of P, denoted by $Rank(P, Cite)$ in this set. Similarly, we compute NID at time $y + 10$ for these highly-cited papers and rank them by NID to compute the rank of P, denoted by $Rank(P, NID)$. If NID is a better measure of the paper's impact, then we expect P to have a better rank (1 being the best outcome, i.e., the top paper) compared to the other papers in the compared set. Figure 8 plots $Rank(P, Cite)$ and $Rank(P, NID)$ for each TOT awardee paper P. We note that in most of the cases (25 out of 40), the ToT papers are the top-ranked papers by both citation count and NID. Interestingly, we also note that in 12 out of 40 cases, the ranks of the ToT awardee papers achieved by NID are lower (better) than the ranks achieved by citation counts. Thus, *the papers judged most influential by the community (by giving TOT award) may not always have the highest citations among all their contemporary papers.* There may be some subjective evaluation criteria that capture the influence a paper has had on the field. The results of this experiment indicate that NID is much better at capturing the influence of a paper – 33 out of 40 times, the ToT paper achieves rank 1 when ranked by NID. The overall Mean Reciprocal Rank (MRR) achieved by NID is 0.8771 compared to an MRR of 0.7712 achieved by the citation count. Thus, we can consider NID as a much better surrogate measure of influence for a scientific article.

7 CONCLUSION

We proposed a novel concept, called 'Influence Dispersion Tree' (IDT) to explore and model the structural information among the followup (citing) papers of a given paper linked through citations. We derive several basic and advanced properties of an IDT to understand their relations with the raw citation count. One striking observation is that with the increase in citation count, the depth of an IDT grows much slower than the breadth. However, as the citation count grows, the IDT of a paper moves closer to its ideal IDT configuration. We further proposed a series of metrics to quantify the notion of influence from IDT. Our proposed metric NID turned out to be superior to the raw citation count – (i) to predict how

[4]Many conferences (e.g., SIGIR) nominate top five most cited papers published in a year for the ToT award, in addition to getting nominations from the community.

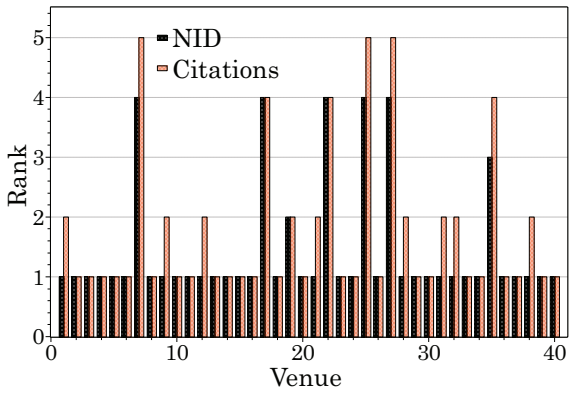


Figure 8: Absolute ranks (based on citation count and NID) of the ToT papers among their contemporaries.

many new citations a paper is going to receive within a certain time window after publication, (ii) to identify and explain why a paper is recognized by its research community (through various prestigious awards such as Test of Time awards) as highly influential among its contemporaries. We conclude that in order to understand the contribution of a source paper to a research field, in addition to the total number of followup papers of a source paper (i.e., citation count), one should also consider how these followup papers are organized among themselves through citations. A paper can be treated as highly influential only when it has enriched a field equally in both vertical (deepening the knowledge further inside the field) and horizontal (allowing the emergence of new sub-fields) directions.

ACKNOWLEDGEMENT

Part of the research was supported by the Ramanujan Fellowship, Early Career Research Award (ECR/2017/001691) (SERB, DST), and the Infosys Centre for AI at IIITD. Dattatreya Mohapatra and Abhishek Maiti were supported by SIGIR travel grants.

REFERENCES

[1] Tomas C Almind and Peter Ingwersen. 1997. Informetric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of documentation* 53, 4 (1997), 404–426.
[2] Frank M Bass. 1969. A new product growth for model consumer durables. *Management science* 15, 5 (1969), 215–227.
[3] Johan Bollen and Herbert Van de Sompel. 2006. Mapping the structure of science through usage. *Scientometrics* 69, 2 (2006), 227–258.
[4] Lutz Bornmann, Rüdiger Mutz, Sven E Hug, and Hans-Dieter Daniel. 2011. A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *Journal of Informetrics* 5, 3 (2011), 346–359.
[5] Tanmoy Chakraborty, Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, and Animesh Mukherjee. 2015. On the categorization of scientific citation profiles in computer science. *Commun. ACM* 58, 9 (2015), 82–90.
[6] Tanmoy Chakraborty and Subrata Nandi. 2018. Universal trajectories of scientific success. *Knowledge and Information Systems* 54, 2 (2018), 487–509.
[7] Tanmoy Chakraborty and Ramasuri Narayanam. 2016. All fingers are not equal: Intensity of references in scientific articles. In *EMNLP*. 1348–1358.
[8] Chaomei Chen. 2018. Cascading Citation Expansion. *CoRR* abs/1806.00089 (2018). http://arxiv.org/abs/1806.00089
[9] Chaomei Chen and Diana Hicks. 2004. Tracing knowledge diffusion. *Scientometrics* 59, 2 (2004), 199–211.
[10] Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can Cascades Be Predicted?. In *WWW*. 925–936.
[11] James Connor. 2011. Google Scholar citations open to all. https://scholar.googleblog.com/2011/11/google-scholar-citations-open-to-all.html. Accessed: April 20, 2019.
[12] Ying Ding, Erjia Yan, Arthur Frazho, and James Caverlee. 2009. PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology* 60, 11 (2009), 2229–2243.
[13] Leo Egghe. 2006. An improvement of the h-index: The g-index. ISSI.
[14] Alan Fersht. 2009. The most influential journals: Impact Factor and Eigenfactor.
[15] Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. 2010. A survey of graph edit distance. *Pattern Analysis and applications* 13, 1 (2010), 113–129.
[16] Eugene Garfield. 1964. " Science Citation Index"-A New Dimension in Indexing. *Science* 144, 3619 (1964), 649–654.
[17] Eugene Garfield. 1972. Citation analysis as a tool in journal evaluation. *Science* 178, 4060 (1972), 471–479.
[18] Eugene Garfield. 2006. The history and meaning of the journal impact factor. *Jama* 295, 1 (2006), 90–93.
[19] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. 2015. The structural virality of online diffusion. *Management Science* 62, 1 (2015), 180–196.
[20] Stefanie Haustein, Isabella Peters, Judit Bar-Ilan, Jason Priem, Hadas Shema, and Jens Terliesner. 2014. Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics* 101, 2 (2014), 1145–1163.
[21] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences* 102, 46 (2005), 16569–16572.
[22] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences* 102, 46 (2005), 16569–16572.
[23] George S Howard. 1983. Research productivity in counseling psychology: An update and generalization study. *Journal of Counseling Psychology* 30, 4 (1983), 600.
[24] Yong Huang, Yi Bu, Ying Ding, and Wei Lu. 2018. Number versus structure: towards citing cascades. *Scientometrics* 117, 3 (2018), 2177–2193.
[25] John PA Ioannidis. 2008. Measuring co-authorship and networking-adjusted scientific impact. *PLoS One* 3, 7 (2008), e2778.
[26] Michael J Kurtz and Johan Bollen. 2011. Usage bibliometrics. *arXiv preprint arXiv:1102.2891* (2011).
[27] Janet Lee, Kristin L Kraus, and William T Couldwell. 2009. Use of the h index in neurosurgery. *Journal of neurosurgery* 111, 2 (2009), 387–392.
[28] Facundo Mémoli. 2011. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics* 11, 4 (2011), 417–487.
[29] Chao Min, Ying Ding, Jiang Li, Yi Bu, Lei Pei, and Jianjun Sun. 2018. Innovation or imitation: The diffusion of citations. *Journal of the Association for Information Science and Technology* 69, 10 (2018), 1271–1282.
[30] Chao Min, Jianjun Sun, and Ying Ding. 2017. Quantifying the evolution of citation cascades. *Proceedings of the Association for Information Science and Technology* 54, 1 (2017), 761–763.
[31] Alex Post, Adam Y Li, Jennifer B Dai, Akbar Y Maniya, Syed Haider, Stanislaw Sobotka, and Tanvir F Choudhri. 2018. c-index and Subindices of the h-index: New Variants of the h-index to Account for Variations in Author Contribution. *Cureus* 10, 5 (2018).
[32] Dinesh Pradhan, Partha Sarathi Paul, Umesh Maheswari, Subrata Nandi, and Tanmoy Chakraborty. 2017. C3-index: a PageRank based multi-faceted metric for authors' performance measurement. *Scientometrics* 110, 1 (01 Jan 2017), 253–273.
[33] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. 2008. Universality of citation distributions: Toward an objective measure of scientific impact. *PNAS* 105, 45 (2008), 17268–17272.
[34] James Ravenscroft, Maria Liakata, Amanda Clare, and Daniel Duma. 2017. Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. *PloS one* 12, 3 (2017), e0173152.
[35] Sidney Redner. 1998. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems* 4, 2 (1998), 131–134.
[36] Andrej A Romanovsky. 2012. Revised h index for biomedical research.
[37] Mayank Singh, Vikas Patidar, Suhansanu Kumar, Tanmoy Chakraborty, Animesh Mukherjee, and Pawan Goyal. 2015. The role of citation context in predicting long-term citation profiles: An experimental study based on a massive bibliographic text dataset. In *CIKM*. ACM, 1271–1280.
[38] Mayank Singh, Rajdeep Sarkar, Pawan Goyal, Animesh Mukherjee, and Soumen Chakrabarti. 2017. Relay-Linking Models for Prominence and Obsolescence in Evolving Networks. In *SIGKDD*. 1077–1086.
[39] Dennis F Thompson, Erin C Callen, and Milap C Nahata. 2009. New indices in scholarship assessment. *American Journal of Pharmaceutical Education* 73, 6 (2009), 111.
[40] Michaël Charles Waumans and Hugues Bersini. 2016. Genealogical trees of scientific papers. *PloS one* 11, 3 (2016), e0150588.
[41] Chun-Ting Zhang. 2009. The e-index, complementing the h-index for excess citations. *PLoS One* 4, 5 (2009), e5429.
[42] Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology* 66, 2 (2015), 408–427.