Independency of Knowledge Diffusion Analyzed by Inverse Citation Networks

Qiaozheng Chi* and Yinghong Ma

Business School, Shandong Normal University, Shandong, 250300, China

*Corresponding author

Abstract. With the development of worldwide knowledge-based economy, structures of knowledge diffusion in scientific research have become extremely complex and dynamic. Properly evaluating the knowledge diffusion would encourage authors to pursue high quality researches. Hence, this paper presents a novel metric of independency of knowledge diffusion (IKD) on the published paper v, defined as the ratio of citation counts of v without its references' to citation counts of v and its references' minus their commons'. Utilizing the inverse citation network formed by published papers in American Physical Society (APS) from 1997 and 2016, the experimental results show that the distributions of IKD are following power law behaviors and the values of IKD are affected by citation counts and involved cooperative institutions. It is reasonable to assess the performances of knowledge diffusion by the metric of IKD.

Keywords: Independency of knowledge diffusion; Inverse citation network; Statistical analysis.

1. Introduction

Evaluating the knowledge innovation is currently a common practice of funding agencies, fellowship evaluations and hiring institutions [1]. Many approaches have been proposed to evaluate the performances of publications, individuals or institutions. The statistical measures, such as the average on the number of citations, the number of papers published per year, or citation counts of the single are some traditional methods to evaluate individuals' performance [2]. Especially, h-index is one of the most popular indicators proposed by Hirsch to quantify both the productivity and the citation impact of the publications for a scientist or a scholar [3]. Many researches were devoted to modify the h-index to improve or optimize the efficiency and the accuracy of it. For example, the g-index [4], the h’-index [5], the hg-index, the hm-index and the hi-index [6,7] and so on. Whatever the h-index or the expanded indexes, the aims of them are to assess the individual's performance not the published papers'. However, the quality of paper cannot be measured or fully captured by citation numbers. For example, citation counts can be easily gamed and manipulated, and the h-index disadvantages early career academics [8]. About 54 false papers were added to Google scholar attributing falsely a total of 325 citations to the author not belonging to the author [9]. The phenomena of "citation by an elite of an elite for an elite" [10] still exists. The citing behaviors such as the co-citations, the institutions are ignored in the pure statistical measurements.

Inspired by the above analysis, we take the citation behaviors into account to measure the knowledge diffusion performance of the paper. On the one hand, the evaluation based on statistical properties of the citations and the publications without the local information of the citing papers and their references can't completely measure the diffusion behaviors of the paper, or lead to deviations. For example, the number of citations of u and its citations v1, v2 are 2 in Fig. 1, while v1 and v2 are much different with their neighborhoods in the three cases. On the other hand, the co-authors in one paper are coming from
different research fields, institutions, and they have different cooperative behaviors. Hence, they often work with a few of cooperators and cite the classical papers, but not always the latest. Therefore, it is interesting to find whether scientific cooperation improves knowledge diffusion or not, whether the size of cooperative team effects knowledge flow full, partly or none.

![Figure 1. Concept illustration. Nodes are papers, the direct links are the inverse citations.](image1)

2. Data

The data set for experiments in this paper is the physical science research papers published in journals of American Physical Society (APS). The reasons for choosing those data are that physical science not only is one of the leading disciplines in natural science, connecting with material science and other high-tech science, but also has multidisciplinary foundations enhancing knowledge diffusion and transferring. The parameters of each paper include the title, journal's name, published date, authors, affiliations, citations and references. There are 191503 papers and 3329319 citations from the year 1997 to 2016. There are two class of research carriers: one is the inverse citation network, the other is the institutions or affiliations. By those two carriers, we evaluate the path independency of knowledge diffusion.

There are 11 dependent time-serial citation networks constructed in this part. Each one includes ten years' citations with two parts: one part is the accumulated citations of the paper's references forward five years before the paper published and backward five years after it published in the data set; and the other part is the accumulated citations of the paper in five years after it published. Hence, each citation network is a ten years accumulation.

3. Defining matrix of IKD

![Figure 2. A concept graph of IKD for a fixed time.](image2)

In this issue, research object is the published papers, the direction is from a published paper (the sender) to the citing paper (the receiver). Here, a network is denoted by \( G=(V, E) \), where
\(V\) and \(E\) are the nodes set and the direct edges set respectively. In this paper, a node in \(V\) is a paper, and the direction from the node \(u\) to the node \(v\) shows that the paper \(u\) is one of references of \(v\). That is, \((u, v)\) is a direct link, and \(v\) is the citation of \(u\). In Fig. 2, for example, the knowledge of the paper \(j\) is studied by its citations \(i, m, f, e\). In other word, the knowledge of paper \(j\) flows to papers \(i, m, f, e\). In the same way, the knowledge of the papers \(i\) and \(j\) are studied by papers \(m, f\), the knowledge of papers \(i\) and \(k\) are studied the same paper \(g\). While, there is some different between papers \(g\) and \(f\) even the two papers all cited \(i\). Papers \(j, i, f\) formed a triangle, while \(j, i, g\) is a path. Paper \(f\) cited the paper \(i\) together with its reference \(j\).

As a result, the independency of the knowledge diffusion (IKD) of a paper \(v\) at a moment is defined as the ratio of citation counts of \(v\) without its references' to citation counts of \(v\) and its references' minus their commons'. It is formed by three sets: one set is papers who cite \(v\), one set is papers who cite references of \(v\), and the other set is the commons of the previous two sets. Supposing the paper \(u\) be the reference of \(v\). At a certain time \(t\), some papers cite the paper \(v\) or its references, denote \(A^\text{out}_v = \{x \mid (v, x) \in E, v, x \in V\}\) and \(B^\text{out}_{u \leftarrow v} = \bigcup_u \{y \mid (u, y) \in E, u, v, y \in V\} = \bigcup_u \bigcup_{u \leftarrow v}\). \(A^\text{out}_v\) be the sets of citations of \(v\) and its references' respectively. The commons neighbor set of the two previous sets is \(A^\text{out}_v \cap B^\text{out}_{u \leftarrow v}\). Then, the independency of knowledge diffusion of a node \(v\) at time \(t\) is defined as Eq. 1 where \(|A|\) is the size of the set \(A\):

\[
IKD_v(t) = \frac{|A^\text{out}_v| - |A^\text{out}_v \cap B^\text{out}_{u \leftarrow v}|}{|A^\text{out}_v| + |B^\text{out}_{u \leftarrow v}| - |A^\text{out}_v \cap B^\text{out}_{u \leftarrow v}|} \tag{1}
\]

In Fig. 2, for example, let \(v=i, u=j\), then \(A^\text{out}_i = \{f, g, h, m\}\) and \(B^\text{out}_{i \leftarrow j} = \{m, f, e, i\}\) respectively, then \(IKD_j(t) = 0.333\) by Eq. 1. Similarity, \(IKD_k(t) = 0.333\) and \(IKD_h(t) = 0.667\) respectively. If \(A^\text{out}_v \cap B^\text{out}_{u \leftarrow v}\) is empty, the sizes of \(A^\text{out}_v\) and \(B^\text{out}_{u \leftarrow v}\) determine the value of \(IKD_v\) completely; Otherwise, if \(A^\text{out}_v \cap B^\text{out}_{u \leftarrow v}\) is not empty, the node in the intersection of \(A^\text{out}_v\) and \(B^\text{out}_{u \leftarrow v}\) forms the triangle structures: \(v\), the reference \(u\) and their intersection citations. Whatever the pattern of \(A^\text{out}_v \cap B^\text{out}_{u \leftarrow v}\) is, the neighbors of the sources and the terminal of the node \(v\) are considered, the metric of \(IKD_v\) displays the local structure.

When the time \(t\) is evolving, the number of citations of a paper cumulates since it published, so \(IKD_v(t)\) is a dynamical index. For a large enough time \(T\), \(IKD_v(t)\) is defined as the biggest value of \(IKD_v(t)\) when \(t \rightarrow T\). That is,

\[
IKD_v = \max_{t \rightarrow T} \{IKD_v(t) \mid 1 \leq t \leq T\} \tag{2}
\]

4. Statistics analysis on IKD

![Figure 3](image_url)

(a) The distribution of citations on papers in citation network cumulated from 1997 to 2016.
(b) The distribution of the values of IKD of the 11 citation networks.
The two panels in Fig. 3 show the distributions of citations and densities of the node's IKD on the 11 citation networks respectively. Fig. 3(a) displays power-law and fat tail properties, shapes in Fig. 3(b) are almost the same except the peaks of them. The values of most papers' IKD are in the interval 0 to 0.25, and some of them near to 1. Some of papers with high value of IKD might be the important papers or comprehensive reviews.

![Figure 3](image-url)

**Figure 3.** The distributions of citations and densities of the node's IKD on the 11 citation networks.

Denote \(<\text{IKD}>\) and \(<\text{C}>\) be the average IKD and the average citations in networks respectively. The values of \(<\text{C}>\) and \(<\text{IKD}>\) are tending to steady in the 11 citation networks, shown as black curves in Fig. 4(a) and Fig. 4(b) respectively. The two blue curves in the two panels of Fig. 4 are the values of \(<\text{C}>\) and \(<\text{IKD}>\) of the top 1% citing papers respectively. The two blue curves excess the blacks which coincide with the instinct that the number of citations of a paper positively correlated with its IKD. The biggest values of \(<\text{IKD}>\) is 0.11 in the year of 2008, shown in Fig. 4(b).

However, Fig. 4 shows many differences between citations and values of IKD. The overall trend of the average citations is increasing slowly, while the values of \(<\text{IKD}>\) is dynamic. The top citing papers display a dynamic increasing behavior in \(<\text{C}>\), but \(<\text{IKD}>\) shows extremely unsteady behavior. This means the independency of knowledge diffusion is influenced not only by citations, but also by the other factors.

![Figure 4](image-url)

**Figure 4.** The average on citations and IKD of all papers and the top 1% citations papers.

In Fig. 5, we divide the published papers into two parts, one is papers with no more than 5 institutions and the other is more than 5. There are two aspects for such partition: on the one hand, there is much diversity of the number of published papers with different number of cooperative institutes, and there is a big slope of the trend. More than 97.5% of the 150344 papers in APS data are cooperated by no more than 5 institutions, and about 2.5% papers (3831) are cooperated more than 5; on the other hand, the cooperation networks formed by the cooperative authors of papers are scale-free networks, and it is one of the social networks with the small world phenomena [11] in which the average path of any two authors is about 5.

![Figure 5](image-url)

**Figure 5.** The relationship of \(<\text{IKD}>\) and the number of institutions. (a) is the case of the number of institutions no more than 5 and (b) is the case of the number of institutions greater than 5.
5. Discussions and conclusions

In this issue, we investigate the visualized structures of citation networks and put forward a novel metric IKD to measure the performance of knowledge diffusion of papers. With the data of APS from 1997 to 2016, experiments are taken on the inversed citation networks. We find that the distributions of IKD are following power law behaviors and the values of IKD are affected by the citation counts and the number of involved institutions in time-window networks. According to the findings, scientific policy makers should take both researchers' outputs and IKD of their achievements into consideration to evaluate the performance of researches. And in order to create innovative achievements, researchers should pay more attention to what they research, to whom they collaborate with and to what resources they could take to use.

In the future work, there are two directions to further investigate on this research. The first is to empirically analyze the structural patterns of citation network and knowledge diffusion. This future work might be dependent upon the quality of data. The second is to identify the performance of different institutions on knowledge diffusion which may contribute to a further insight on evaluating the qualities of institutions. However, the relationships between knowledge diffusion and innovation are still deserved to be discussed.

References

[1] C. O. S. Sorzano, J. Vargas, G. Caffarena-Ferna´ndez, A. Iriarte, Comparing scientific performance among equals, Scientometrics 101(3) (2014) 1–15.
[2] W. Glanzel, On the opportunities and limitations of the h-index, Vision Research 35(10) (2006) 1393–1399.
[3] J. E. Hirsch, An index to quantify an individual’s scientific research output, PNAS 102(46) (2005) 16569–16572.
[4] Egghe L. An improvement of the h-index: the g-index, ISSI newsletter 2(1) (2006) 8-9.
[5] Hirsch, Jorge E. ha: An index to quantify an individual’s scientific leadership. Scientometrics 118(2) (2019) 673-686.
[6] M. Schreiber, A modification of the h-index: The hm-index accounts for multi-authored manuscripts, Journal of Informetrics 2(3) (2008) 211–216.
[7] M. Schreiber, A skeptical view on the hirsch index and its predictive power, Physica Scripta 93(10) (2018) 102501.
[8] G. Thorsten, Academic sell-out: how an obsession with metrics and rankings is damaging academia, Journal of Marketing for Higher Education 24(2) (2014) 165–177.
[9] A. T. S. Jaime, The google scholar h-index: useful but burdensome metric, Scientometrics 117 (2018) 631–635.
[10] M. Stuart, K. Jacqueline, Counting footnotes: Citability in management studies, Scandinavian Journal of Management 26(2) (2010) 189–203.
[11] E Caixxell, Science citation index. Effectiveness in local articles in the anaesthesics field: "perturbation of iontransport", British journal of anaesthesia 43(8) (1971) 814–818.