Number versus Structure: Towards Citing Cascades

Yong Huang †

Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan, Hubei, China

Yi Bu †

Center for Complex Networks and Systems Research, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN., U.S.A.

Ying Ding

School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN., U.S.A.

School of Information Management, Wuhan University, Wuhan, Hubei, China

School of Management, Tianjin Normal University, Tianjin, China

School of Management, Jilin University, Changchun, Jilin, China

Wei Lu *

Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan, Hubei, China

†: Equal contribution.

*: **Correspondence concerning this article should be addressed to Dr. Wei Lu**, Email: weilu@whu.edu.cn.

Number versus Structure: Towards Citing Cascades

Abstract: This paper proposes a novel concept of the citing cascade, defined as a network comprising citing relationships between a paper and its citing paper, as well as those among its citing papers. Compared with citation counts using a single number, citing cascades reveal the structural information of citation networks of a scientific publication and thus help us better understand the citation impact of a scientific publication (called the owner of the citing cascade). We then define and elaborate on several basic and advanced properties of citing cascades. By employing computer science publication records in the Microsoft Academic Graph dataset, we found that cascade size, edge count, in-degree, and out-degree all follow typical power law distributions with various exponential parameters (α). In addition, cascade depth is observed to follow an exponential distribution. We also examine the relation between citation count of the owner and some advanced properties that we defined. Many related future studies are also illustrated at the end of this paper.

Keywords: citing cascade; citation count; citation impact; scientometrics.

INTRODUCTION

Citation counts have been adopted as a dominant indicator in research evaluation for decades (Waltman, 2016). Yet, scientometricians have never stopped improving on this indicator, such as normalizing the raw citation count from various perspectives (e.g., Radicchi, Fortunato, & Castellano, 2008; Waltman & Van Eck, 2015), implementing PageRank-related strategies considering the citing publications' citation impact (e.g., Ding, Yan, Frazho, & Caverlee, 2009; Waltman & Yan, 2014), employing full-text data to distinguish citations with different occurrence time or location in the full text (e.g., Ding, Liu, Guo, & Cronin, 2013; Zhao, Cappello, & Johnston, 2017), and taking into the consideration citation networks (e.g., Kuhn, Perc, & Helbing, 2014; Perc, 2010,

2013).

Nevertheless, these approaches have dealt with the citation impact of a scientific publication as a single number although some of them considered network-oriented issues. Yet, seldom of them considered the citing relationships between its citing publications except our previous work (Huang, Bu, Ding, & Lu, 2018). Following this work, this present paper explores the structural information among these relationships, since they are of importance in understanding and quantifying the citation impact of a publication. To this end, we propose a novel network structure, namely citing cascades[1]. Different from citation count that uses a simple number, a publication's citing cascade is a citation network containing the citing relationships between the publication and its citing publications, and citing relationships within its citing publications. Figure 1 is an example of a citing cascade, in which citing relationships between publication A and its citing publications, as well as citing relationships among A's citing publications, are included. Note that citing relationships pointing to A are shown in solid lines, while those not dotted lines.

[1] In Huang *et al.* (2018), such networks are named as "ego-centered citation networks."

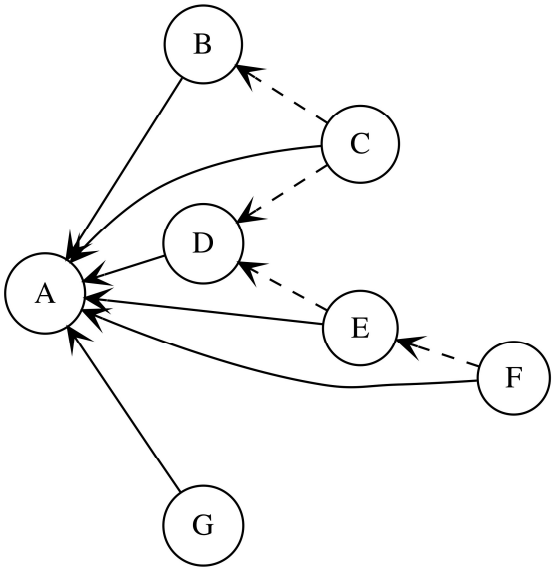

Figure 1. Illustration of a citing cascade. Each node represents a publication, where publication A is the owner of this citing cascade (Huang *et al.*, 2018). Each edge indicates a citing relationship from the source node (citing publication) to the target node (cited publication). Citing relationships pointing to A are shown in solid lines, while those not dotted lines.

A cascade is defined as a process "whereby something, typically information or knowledge, is successively passed on[2]." When information is conveyed and passed on, the cascade is called an *information cascade*. The main reason why the network structure in Figure 1 is named as a citing cascade is that a *citing cascade* is a specific type of information cascade in which citing relationships are successively transmitted. Defining citing cascades help us understand the citing relationships (structural details) between a publication's citing publications and paint a more nuanced pictures on the citation impact of the raw publication.

The outline of this article is as follows. Related work concerning information cascades is provided in Section 2. The definition and several properties of citing cascades are proposed in Section 3. The dataset used in this paper and the results of the empirical

[2] https://en.oxforddictionaries.com/definition/cascade

studies are illustrated in Section 4. Finally, the conclusion and future work are presented in Sections 5 and 6.

INFORMATION CASCADES

The definition of information cascade is unclear and disparate in various disciplines. The earliest definition of information cascades derives from the field of sociology. For instance, Bikhchandani, Hirshleifer, and Welch (1992) defined information cascade as a way to interpret herding behavior (Trueman, 1994) (definition I) in which each decision-maker looks at decisions by considering previous decision-makers in spite of their own information. Using several detailed circumstances (e.g., fashion industry) to show how information cascades are able to interpret them, they pointed out the fragility and idiosyncratically of conformist behavior because of information cascades that "start readily on the basis of even a small amount of information" (p. 1016). In their paper, they also argued that (p. 1016):

"Conformity often appears spontaneously without any obvious punishment of the deviators. Informational cascades can explain how such social conventions and norms arise, are maintained, and change."

However, later researchers outside of sociology expanded on this early definition and demarcated information cascades under their own context. In the present paper, cascades with information successively passed on is termed *information cascades* (definition II), which has been adopted by numerous researchers (e.g., Anderson & Holt, 1997; Watts, 2002). An information cascade under definition II could be articulated by the following three aspects:

(1) ***What is diffused***: In general, "information" is diffused in an information cascade. However, in some studies, people care more about a specific item that is conveyed. Alvarez, Garcia, Moreno, and Schweitzer (2015), for instance, investigated

microblogging users' sentiments diffusing in an information cascade. Anderson, Huttenlocher, Kleinberg, Leskovec, and Tiwari (2015) examined how "invitations" are diffused among LinkedIn users' network. The entity that is diffused in a cascade is defined as its subject.

(2) ***How the subject is diffused***: People make decisions under the effects of their surroundings, and can "cascade" their decisions to others. Such kind of behavior is termed *cascading behavior* (or in some literature, *cascade behavior*). Apparently, cascading behavior is different from that purely decided by individual reasoning (Leskovec & Singh, 2005). Two effects are highlighted in cascading behavior: (a) people's behavior is influenced by their environment; and (b) people's behavior also influences their surroundings.

(3) ***What the results/outputs are***: The structure or paths of the diffusion of the subject can constitute a tree, graph, or network. The detailed name of a cascade is based on what is conveyed/diffused. We know that, generally, "information" is diffused, and this is thus mostly termed *information cascades*. In the cases of Alvarez *et al.* (2015) and Anderson *et al.* (2015) mentioned in (1), the cascades are named as *sentiment* and *invitation cascades*, respectively. Cascades defined in this paper are termed *citing cascades* because citing relationships are conveyed as defined in Figure 1.

In summary, an information cascade is a concept that differs in other fields compared with that in sociology. Researches from various areas have focused on different aspects of the above three factors.

The fields of management exhibit great interest in how different factors or strategies affect information diffusion in information cascades and how they influence sales volume by implementing regression analyses and/or game theories (Kleinberg, 2007). Duan, Gu, and Whinston (2005) measured herding behavior by using the previous number of downloads of a certain software recorded by the CNETD dataset and found that herding behavior in this process is indeed strong. However, they concluded that

herding is not significantly influenced by the provision of professional product reviews or user reviews. Walden and Browne (2002) identified the important role that information cascades play in the scenarios of users' adopting technologies. Specifically, they argued that the process of users to adopt new technologies heavily depends on their prior information. To test this, they regarded information cascade as a process of passing by "signals" from others to a given individual and built an (quantitative) operational model in which the "correct cascade" (herding and making a correct decision) and "incorrect cascade" are both pre-defined. Leskovec and Singh (2005) examined a recommendation network generated from the online seller Amazon, and concluded that a customer receiving more than one recommendation of a certain commodity is more willing to purchase the commodity, which constitutes a typical reflection of an information cascade. Some management scientists also paid attention to social governance through targeting a specific event, such as González-Bailón, Borge-Holthoefer, and Moreno (2013) who built a network based on the data of Twitter users' message activity prior to the 2011 Election Day (May 22). They found that social network dynamics facilitate coordination—in this case, social movements—by triggering information cascades that can potentially reach a large number of people in a short period of time.

Physicists are interested in cascading behavior, mainly exploring the mechanism of this process by presenting quantitative models from a more theoretical perspective. In this area, mainly two branches of study exist. The first presents a certain model and simulates different types of networks by using their proposed model (e.g., Anderson & Holt, 1997; Galstyan & Cohen, 2007; Hisakado & Mori, 2015; Watts, 2002). For instance, threshold models are a set of models that are the most commonly used in this branch of studies. These models assume that nodes (such as persons in an information cascade) display "inertia" in switching states, but once their personal threshold is reached, "the action of even a single neighbor can tip them from one state to another"

(Watts, 2002, p. 5767). Hisakado and Mori (2016) is a good example of using the threshold model to understand information cascades in which public perceptions are conveyed. Specifically, they considered information cascades as a vote process and defined two types of voters, namely independents and herders, in a mathematical way; independents tend to vote according to their fundamental issues, while herders based on the previous number of votes. The empirical studies conducted on random graphs, scale-free graphs generated by the Barabási-Albert (BA) model (Barabási & Albert, 1999), and graphs from fitness models (Bianconi & Barabási, 2001) showed that there are only limited effects of hubs in cascade networks on voters' perceptions. Some other studies exercised independent cascade models (e.g., Wang, Chen, & Wang, 2012a) and Markov chain models (e.g., Li, Ma, Guo, & Mei, 2017; Wang, Scaglione, & Thomas, 2012b). The second branch investigates cascade failure, which is defined as the subsequent failure of a part of a certain network that is induced by other part(s) of the network (e.g., Wang *et al.*, 2012b). Lai, Motter, and Nishikawa (2004) demonstrated that scale-free networks generated by the BA model tend to be more sensitive to short-range, rather than long-range, attacks. Similarly, Buldyrev, Parshani, Paul, Stanley, and Havlin (2010) understood interdependent networks, which means that one network's normal functioning is dependent on another. To do this, they built a single network and several interdependent networks and found that a broader degree distribution of the nodes in interdependent networks correlates to a higher vulnerability on random failure, compared with the single network.

Nevertheless, computer scientists are interested in the final structure and characteristics of information cascades after they have been formed from a network perspective. From macro- and meso-level perspectives, Baños, Borge-Holthoefer, and Moreno (2013) constructed an information cascade using Twitter data and explored several of its properties, including degree (distribution), coreness, average path length, and depth; communities were also detected and analyzed in the cascade. In another study providing

a temporally sentiment analysis, indicators were measured, including cascade size and centrality (Alvarez *et al.*, 2015). Yet, micro-level analyses were missing. Some network phenomena were detected in information cascades. Liben-Nowell and Kleinberg (2008), for instance, built an Internet chain letter network and observed a small-world property; they identified a "narrow but very deep tree-like pattern" in the network (p. 4633). Yet, Golub and Jackson (2010) argued that such pattern might subject to change if one uses a different dataset. Additionally, Anderson *et al.* (2015) studied the homophily in a diffusion process within the LinkedIn invitation cascade; particularly, geography- and industry-related factors are found to have significant homophily effects in the formation of the cascades. Similar research in this stream established empirical studies in viral marketing (Leskovec, Adamic, & Huberman, 2007a), product recommendations (Leskovec, Singh, & Kleinberg, 2006), rumor dissemination (Kostka, Oswald, & Wattenofer, 2008), and social media (Bakshy, Hofman, Mason, & Watts, 2011; Cha, Benevenuto, Ahn, & Gummadi, 2012; Cheng *et al.*, 2014; Cui *et al.*, 2013; Leskovec, McGlohon, Faloutsos, Glance, & Hurst, 2007b; Sun, Rosenn, Marlow, and Lento, 2009; Yu & Fei, 2009).

Overall, we can find that the major concerns of research in management science, physics, and computer science correspond to the aforementioned three aspects: (a) what is diffused; (b) how the subject of a cascade is diffused; and (c) what are the results/outputs. Although having identified and quantified interesting phenomena in information cascades, many of these studies simply concentrated on global scenario but ignored local network patterns of a certain (part of) information cascade. Needless to say, there is also few papers disentangling the joint effects of global versus local patterns, as pointed out by Borge-Holthoefer *et al.* (2013), in a systematical way. Even in the research context of science of science, most network-based studies on citation impacts prefer to use PageRank-related methods to investigate global- but not local-level patterns. To this end, we propose citing cascades that can assist us to better

understand the structural information of a publication's citation network and, therefore, to learn more in-depth about its citation impact.

METHODOLOGY

Definition of citing cascades

A citing cascade is essentially a network. From the perspective of network vertices, the citing cascade of a publication p_0 includes all of the citing publications of p_0 as well as p_0 itself. From the perspective of network edges, it contains all citing relationships from the citing publications of p_0 to p_0 (named as direct citing relationships), plus citing relationships among p_0's citing papers, which are termed indirect citing relationships. A citing cascade can be represented as a directed graph G (V, E) in which each vertex $v \in V$ shows a scientific publication, and every directed edge $e(u, v) \in E$ represents that publication u has ever cited publication v. In a citing cascade, the only vertex whose out-degree equals to zero is named as the ***owner*** of the citing cascade. Those vertices other than the owner, representing the citing publications of the owner, are named as ***endorsers***, as these publications have endorsed (cited) the owner (Ding, 2011). In the citing cascade example in Figure 1, A is the owner, while B, C, D, E, F, and G are all endorsers.

There are three types of endorsers in citing cascades according to the structural information of a citing cascade:

(1) Endorsers having been cited by at least one of the other endorsers, such as B, D, and E, are named as ***connectors***, because in citing cascades they connect other endorsers and the owner. The number of connectors in a citing cascade is annotated as $|V_c|$.

(2) Endorsers that have cited other endorsers, such as C, E, and F, are termed ***late endorsers***, since they were published after the corresponding connectors had been

published. This term is borrowed from the field of innovation diffusion (Brancheau & Wetherbe, 1990). The number of late endorsers in a citing cascade is annotated as $|V_{le}|$. Note that a late endorser might also be a connector at the same time, such as E in Figure 1.

(3) Endorsers that have no indirect citing relationships, regardless of citing others or being cited, with other endorsers, such as G in Figure 1, are named as ***isolate endorsers***. The number of connectors in a citing cascade is annotated as $|V_i|$. Therefore, $|V| - 1 \leq |V_c| + |V_{le}| + |V_i|$.

One of the differences among these three types of endorsers lies in their in- and out-degrees (detailed in the section of "Basic properties of citing cascades").

In summary, we classify all vertices in citing cascades into two types, owner and endorser, as seen in Figure 2. The latter consists of connectors, late endorsers, and isolate endorsers (isolate endorsers are indicated by the grey area); connectors and late endorsers are not always mutually exclusive. Regarding edges, a citing cascade contains direct and indirect citing relationships, shown as solid and dotted lines, respectively, in Figure 1.

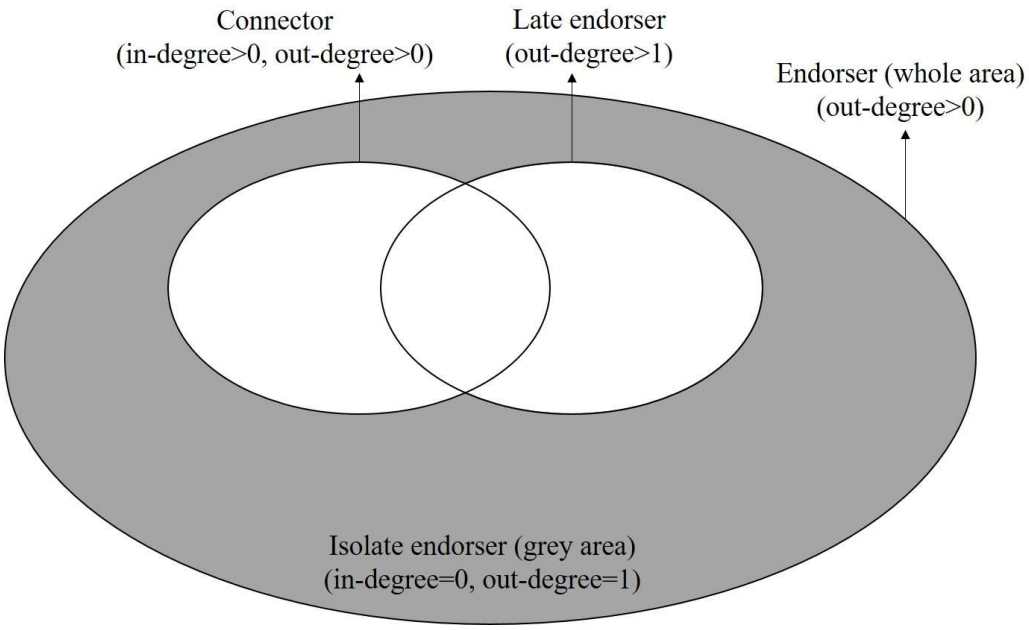

Figure 2. Types of endorsers in a citing cascade.

Information cascades researched in other social networks (e.g., Bakshy *et al.*, 2011) include vertices (followers, essentially endorsers in citing cascades) not linked to the initial vertex (owner in citing cascades). However, our defined citing cascades do not include these vertices, i.e., endorsers' endorsers that have never cited the owners are excluded.

Basic properties of citing cascades

A citing cascade is *de facto* a network. We here make an analogy and define five basic properties of citing cascades based on common network properties.

The ***cascade size*** illustrated in a citing cascade is defined as the number of endorsers in a citing cascade, which equals to the ***citation count*** of the owner, $|V|-1$, where $|V|$ refers to the total number of vertices in a citing cascade. For instance, the size of the citing cascade in Figure 1 is six.

The ***edge count*** is the total number of edges in a citing cascade, represented as $|E|$,

including both direct and indirect citing relationships of the owner. The number of direct citing relationships is represented as $|E_D|$, while that of indirect citing relationships is $|E_I|$; obviously, $|E| = |E_D| + |E_I|$. The edge count in Figure 1 equals to 10, six of which are direct citing relationships and the remaining four are indirect citing relationships.

The ***cascade depth*** measures the length of the longest directed path from any endorser to the owner in a citing cascade. The longest directed path in Figure 1 is $[F \to E, E \to D, D \to A]$; therefore, the depth of this citing cascade is three.

Since a citing cascade is a directed graph, the degree of a vertex is measured as ***in-degree*** (the number of edges *linked to* it) and ***out-degree*** (the number of edges *from* it). The in-degree of vertex v is represented as $deg^-(v)$ and its out-degree $deg^+(v)$. If $deg^+(v_0) = 0$, the vertex v_0 is the owner of the citing cascade. If $deg^+(v_0) > 0$ and $deg^-(v_0) > 0$, the vertex v_0 should be a connector. If a vertex v_0 is an endorser but not a connector (i.e., a late or isolate endorser), $deg^-(v_0) = 0$. In the citing cascade shown in Figure 1, for the owner A, its in-degree equals to $deg^-(A) = |V| - 1$, which is equal to citation count, while its out-degree is zero. For C, F, and G, their in-degrees are all zero. The in-degree of the connector B, $deg^-(B)$, is one and the out-degree, $deg^+(B)$, is one; the in-degree of the later adopter C, $deg^-(C)$, is zero and the out-degree is $deg^+(C)$ is three. The relationships between different endorsers, as well as their in- and out-degrees, are presented in Figure 2, in which one can see that all endorsers have out-degrees larger than zero, because they all at least cite the owner. The out-degree of isolate endorsers must be one, as they purely cite the owner. For late endorsers, they cite not only the owner but also certain connector(s), leading their out-degrees to be larger than one. Regarding connectors, their out-degrees could be one, if the connectors are not late endorsers, or larger than one if they are.

Advanced properties of citing cascades

In addition to basic properties analogous to general network attributes, we also define several properties that are characteristic of citing cascades. Firstly, we define the percentage of connectors out of all endorsers, $P(c)$, as:

$$P(c) = \frac{|V_c|}{|V|-1} \quad (1)$$

In the citing cascade shown in Figure 1, $P(c) = \frac{3}{6} = 0.5$. Mathematically, $P(c)$ ranges from $\frac{1}{|E|}$ to $(1 - \frac{1}{|E|})$ if not equivalent to zero.

Similarly, we can define the percentage of late endorsers among all endorsers, $P(le)$, as:

$$P(le) = \frac{|V_{le}|}{|V|-1} \quad (2)$$

We know that all late endorsers' out-degrees are larger than one, thus $P(le) = P(deg^+(v) > 1)$. In the citing cascade presented in Figure 1, there are three endorsers whose out-degrees are greater than one, i.e., C, E, and F; therefore, $P(le) = \frac{3}{6} = 0.5$. Similar to $P(c)$, the range of $P(le)$ is $\{0\} \cup [\frac{1}{|E|}, (1 - \frac{1}{|E|})]$.

We also define the ratio between the numbers of late adopters and connectors to calculate the average number late endorsers linked with a connector in a citing cascade ($ANLEC$). Mathematically, $ANLEC$ is defined as:

$$ANLEC = \frac{|V_{le}|}{|V_c|} \quad (3)$$

In Figure 2, for instance, since we have three late endorsers (C, E, and F) and three connectors (B, D, and E), $ANLEC = 1$. A greater $ANLEC$ indicates a greater number of late endorsers of the given owner linked by the connectors.

Although a connector might be cited many times, not all of its citing publications cite the owner. In other words, some of its citing publications are included, but others are

excluded, from the citing cascade. We therefore define the conversion rate of a connector i, CR_i, as the percentage of citing publications in the citing cascade among all of i's citing publications. In practice, CR_i is calculated as:

$$CR_i = \frac{|V_{le}|_i}{cc_i} \quad (4)$$

where $|V_{le}|_i$ is the number of late endorsers connected by the connector i; and cc_i is the citation count of i recorded in our dataset. For instance, connector B, D and E have been cited five, five, and two times in the whole corpus, and therefore, the converting rate for these connectors are $\frac{1}{5}$, $\frac{2}{5}$, and $\frac{1}{2}$, respectively. Note that in one given citing cascade, connectors might have various conversion rates; meanwhile, a connector might have different conversion rates in citing cascades of distinct owners. In a given citing cascade, the average conversion rate (ACR) is:

$$ACR = \frac{\sum_{i \in V} CR_i}{|V_c|} \quad (5)$$

Both $ANLEC$ and ACR measure how many late endorsers of the owner are linked to the connectors. The difference is that the latter takes into account all of a connector's citing publications (i.e., global citations), while the former simply considers its citing publications within a citing cascade (late endorsers that have cited this connector, i.e., local citations). The two properties complement each other and assist us to better understand the structures of citing cascades.

RESULTS AND DISCUSSION

Data

All publications labelled as "Computer Science" in the Microsoft Academic Graph (MAG) dataset (Sinha *et al.*, 2015) are utilized in this study, annotated as MAG-CS. This results in a total of 5,249,815 publications in the dataset, among which there are 2,429,009 papers that have received at least one citation with no citation loops or circles. We then built a citing cascade for each of these papers, containing the citing

relationships between a publication and its citing publications, as well as those among its citing publications. Additional details on the descriptive statistics on the dataset that we used can be found in our previous work (Huang *et al.*, 2018).

Distributions of basic properties in citing cascades

We plot the probability distribution (PD) of five basic properties in citing cascades within the MAG-CS dataset, as shown in Figure 3, in which the vertical axes represent the probability (i.e., percentage) that a variable equals to the corresponding horizontal axis value.

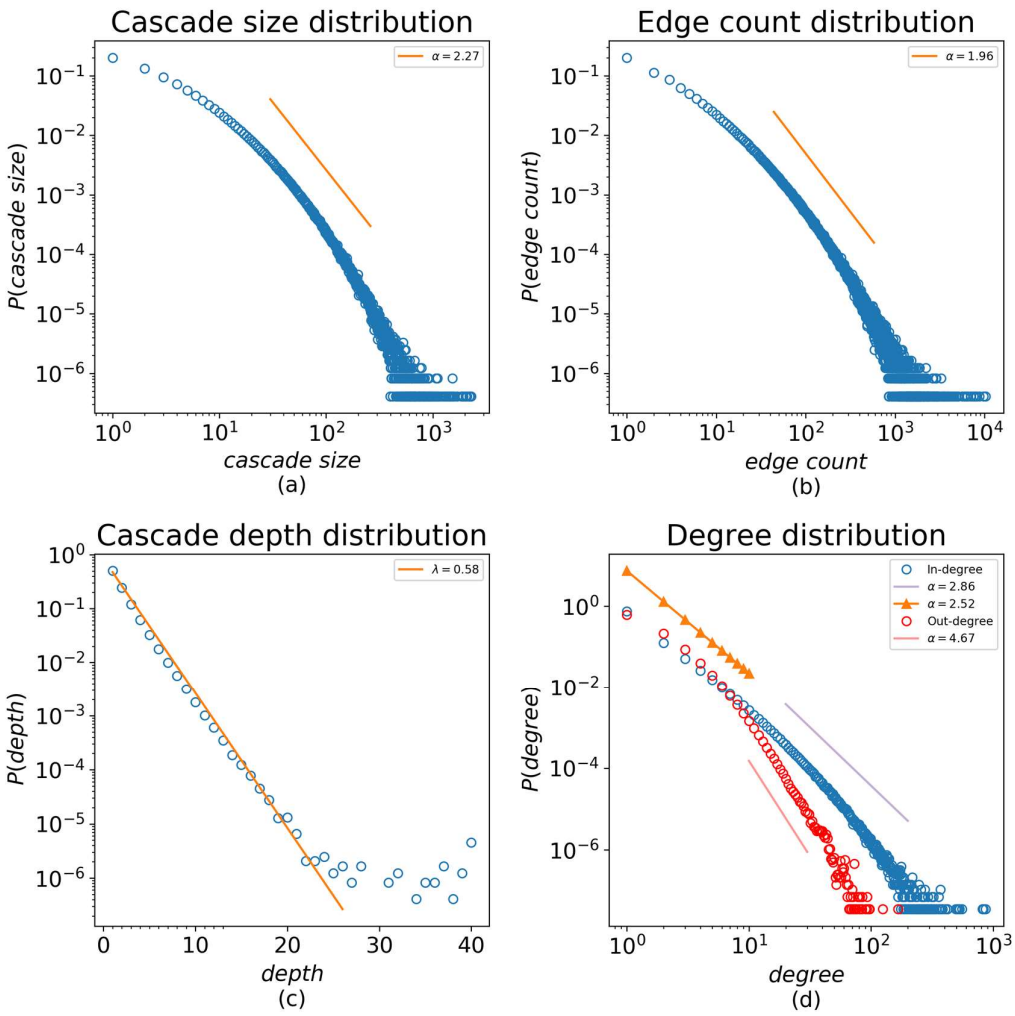

Figure 3. Probability distributions (PD) of basic properties in citing cascades: cascade size (a), edge count (b), cascade depth (c), and in- and out-degree (d). In panels (d), all in-degree values are normalized by adding one to prevent the calculation of $\boldsymbol{lg0}$.

From Figures 3(a) and 3(b), it can be seen that the dots follow a downward trend, meaning that the dots with small values dominate in quantity, and those with large values are fewer. Specifically, approximately 2.4% of all cascades have a cascade size of 10, but the possibility of selecting a citing cascade whose size is 100 is only ~0.02%; approximately 0.085% of the cascades feature an edge size of 20. We also find that the fitted curve is a right-skewed straight line on log-log scales, indicating that the cascade size and edge count both follow power law distributions, in general. Mathematically, a variable x follows a power law distribution if its PD $p(x) = kx^{-\alpha}$, where k is a

constant and α is known as the exponential parameter (Clauset, Shalizi, & Newman, 2009). The red straight lines indicate the power law fitted lines, with α equal to 2.27 and 1.96, respectively. The estimated values of α can reveal important properties of how the mean and variance of the distribution scale with system size (Alvarez *et al.*, 2015), which in our case is the amount of papers in the MAG-CS dataset. For example, $\alpha \leq 2$ implies that both the mean and the variance of x increase with the size of the sample, but $\alpha > 2$ indicates that those would not scale with the system size (Newman, 2005). From our empirical results, it can be seen that the property of cascade size is fitted with $\alpha > 2$, while edge size $\alpha \leq 2$. These mean that the expected cascade size would not be greater with a larger amount of publications in the dataset. Nonetheless, this is not the case for edge size, in which the expected edge size would increase as the dataset size increases.

Cascade depth, as a commonly used structural property, reveals the depth of influence while the cascade size indicates the width of influence. In Figure 3(c), one can see that approximately 0.1% of all citing cascades in MAG-CS have a depth of 10, meaning that the longest path in these cascades is 10. A cascade (assume that A is its owner) whose depth is 10 means that there is at least one citing publication of the owner (annotated as J) that has cited another citing publication of the owner (I); I also cited H, …, C cited B, and B cited the owner A; all $B - J$ cited the owner A, as shown in Figure 4. A cascade like this reveals an extremely large influence of the owner on its related field. Meanwhile, the scatters appear to be fitted by the solid line under an exponential distribution $p_{depth} \sim \lambda e^{-\lambda x}$ upon a semi-logarithmic scale, where $\lambda = 0.58$. According to the mathematical properties of exponential distributions, the mean value of the cascade depth equals to $\frac{1}{0.58} \approx 1.72$. This is reasonable because there are many publications that are only cited once; in these cases, the cascade depth is equivalent to one. Regarding in- and out-degrees, we adopt blue dots to represent in-degree and red out-degree in Figure 3(d), in which we find that the indicators of in- and out-degree

decay follow typical power law distributions. In order to prevent the calculation of $lg0$, in practice, the values of in-degree in the horizontal axes are normalized by adding one to the actual in-degree value. From the figure, we observe similar distributions for in- and out-degrees. Dots on the in- and out-degree curves coincide with each other when degree values are smaller than 10 (following a power-law distribution with an exponent parameter $\alpha = 2.52$, shown as orange triangles). The out-degree curve decreases much faster, decaying as a power law distribution with $\alpha = 5.33$, than in-degree that is characterized by a power law distribution with $\alpha = 2.98$.

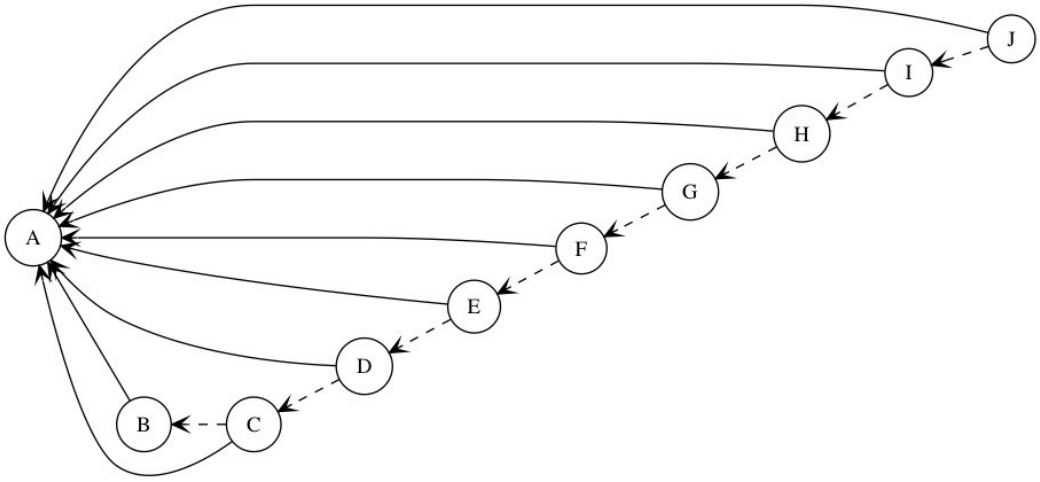

Figure 4. A simplified example of a citing cascade with depth=10.

Correlation between citation count of the owner and advanced properties of citing cascades

For each advanced property defined in the "Advanced properties of citing cascades" section, we plot a heat scatter plot in Figure 5. In each subfigure, the horizontal axis represents the citation count of the owner in a citing cascade, while the vertical axis the corresponding property (e.g., in Figure 5(c), the vertical axis refers to the percentage of late endorsers). The color indicates the number of citing cascades whose owners' citation counts and a certain property equal to the corresponding values shown in two

axes; the color follows the bar to the right of the figure. For instance, the number of citing cascades that contain 50% connectors out of all endorsers is 3,150 (10^3 scale), if we constrain the citation count of the cascade owner as 10 in the MAG-CS dataset. In addition to the scatters in Figure 5, several other curves are displayed. The pink lines that usually exhibit many fluctuations are composed of the mean values of a certain property at each citation count. For example, in Figure 5(b), the average value of ACR when the citation count equals to 40 is 4.12%. For better visualization, we also plot red solid curves that are the smoothed pink lines by employing an algorithm called Locally Weighted Scatterplot Smoothing (Cleveland, 1979). Meanwhile, in Figures 5(c) and 5(d), the grey curves show the upper and lower bound of the corresponding properties, except those whose $P(c)$ or $P(le)$ equals to zero.

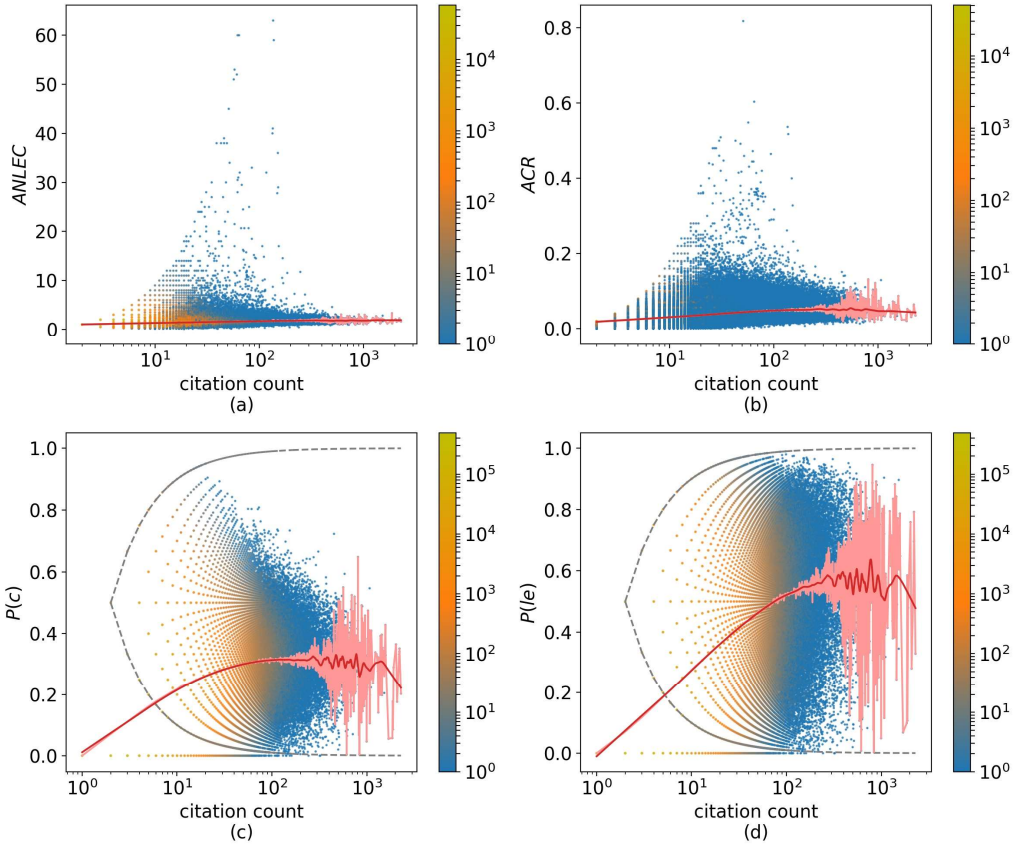

Figure 5. The relation between the citation count of the owner and advanced properties of citing

cascades in MAG-CS: (a) average number of late endorsers linked with a connector ($\boldsymbol{ANLEC}$); (b) average conversion rate ($\boldsymbol{ACR}$); (c) the percentage of connectors ($\boldsymbol{P(c)}$); and (d) the percentage of late endorsers ($\boldsymbol{P(le)}$).

The dots in the left of Figure 5(a) are orange while those in the right tend to be blue, indicating that there are more publications with a lower number of citations than those with greater citation counts. For example, there are 40 citing cascades in which the owners' citation count is 20 and an average of five late endorsers connected to one connector, but the corresponding number for owners being cited 100 times and $ANLEC$ five is only one. Meanwhile, we observe in Figures 5(a) and 5(b) some blue dots with medium horizontal values but relatively higher vertical values, which reveals that in citing cascades of medium-cited publications, it is likely to contain more late endorsers than connectors and feature greater conversion rates than others. In Figure 5(b), it can be seen that that the number of orange dots is very limited, and most of them are located in the left part of the figure (fewer than 20 citations). This phenomenon implies the rareness of high conversion-rate connectors in citing cascades with highly-cited owners.

From the perspective of the red line indicating smoothed average values in both Figures 5(a) and 5(b), we do not detect obvious changes when the citation count increases. We know that the property of $ANLEC$ illustrates the average number of late endorsers linked by a connector. Hence, the invariability of this property indicates that a single connector, regardless of which citing cascades of publications with different numbers of citations, has a relatively similar number of linked late endorsers of the owner.

Different from $ANLEC$ that focuses on local scenarios, the conversion rate measures the degree to which a connector is linked by late endorsers of the owner, taking into account all of its citing publications (including those out of a certain citing cascade). The flat red line in Figure 5(b) demonstrates an unchanged conversion rate for connectors in citing cascades with differently-cited owners. Combined with our finding in Figure 5(a), it can be determined that, as owners' citation counts increase, the number

of late endorsers linked by late endorsers of the owner does not change apparently, regardless of considering all connectors' citing publications in the dataset (global citation) or simply the citing publications within the citing cascade (local citation in the cascade).

The heat scatter plots in Figures 5(c) and 5(d) exhibit similar patterns, in which orange dots are dominant on the left and blue ones on the right. However, the maximum values of $p(c)$ first increase and then decrease, while those of $p(le)$ first increase and then remain unchanged. From the perspective of smoothed average value curves, although we observe some fluctuations when owners' citation count is great, generally both figures first show an upward trend, as the citation counts of the owner increase. This finding indicates that lowly-cited publications contain a smaller percentage of connectors and late endorsers in their citing cascade than highly-cited publications. Therefore, citing cascades of lowly-cited publications include more ratios of isolate endorsers, such as publication G in Figure 1. When the number of citations increases, we find that publications' citing cascades tend to have more connectors and late endorsers. However, both of the figures exhibit a flat trend after increasing, which means that after a publication has become highly-cited, $p(c)$ and $p(le)$ do not change obviously as the owners receive more citations. Specifically, the percentages of connectors and late endorsers remain around 0.3 and 0.5, respectively, among the number of all endorsers. We even observe a slight decreasing trend, although nonobvious, in Figures 5(c) and 5(d), when the citation count of the owner is larger than 500. One possible explanation for this is that a highly-cited publication has achieved a relatively sufficient amount of awareness among researchers due the Matthew Effect (Merton, 1968). For example, by default highly-cited related publications are prioritized when scholarly databases, such as Google Scholar, display the retrieval results, and the influence of the connectors on establishing more connections to the owner might be weakened, or at least not be as effective as previously when the

publication was lowly-cited. The findings from Figures 5(c) and 5(d) confirm our previous findings (Huang *et al.*, 2018), in which the number of citing relationships between citing publications increases when lowly- and medium-cited publications receive more citations. Nevertheless, it does not increase significantly as citation counts of highly-cited publication accumulate.

SUMMARY

This paper proposes a novel concept of a citing cascade, which is defined as a network comprising citing relationships between a paper and its citing paper, as well as those among its citing papers. The motivation of defining citing cascades aims to understand the structural information among a publication's citing publications instead of purely using the citation count (a number) as an indicator. We therefore define and elaborate on several basic and advanced properties of citing cascades to understand their patterns beyond citation counts. Several points from our earlier discussions are worth examining in greater detail. By employing the computer science publications records in the Microsoft Academic Graph dataset, we found that cascade size, edge count, in-degree, and out-degree all follow typical power law distributions with various exponential parameters (α). In addition, cascade depth is observed to follow an exponential distribution. We also examined the relation between citation count of the owner and the advanced properties that we defined. Results show that neither the average number of late endorsers connected to a connector ($ANLEC$) nor the average conversion rate (ACR) increases as the number of citations of the owner increases, but the percentages of both connectors and late endorsers first increase and then remain constant.

LIMITATIONS AND FUTURE WORK

If we return to the citing cascade itself, one can find that the definition detailed in this paper did not include the citing relationships between the owner's citing publications

and its citing publications' citing publications. The reason why they were excluded is that the present research simply demonstrates the citation impact of the owner but not further research questions. A citing cascade consisting of the owner, its endorsers, and its endorsers' endorsers should be termed as a *second-order citing cascade* and could be researched in follow-up studies. Each citing publication as an endorser might not only constitute the "spreader" or "adopter" of the owner's idea, but also an "initiator" that would also bring new and innovative inspirations. The initial motivation of an endorser to cite the owner is not likely to assist the owner to disseminate citations more widely, but to help demonstrate the endorser's own ideas. The topics between the owner, its endorsers, its endorsers' endorsers (second-order endorsers), and its endorsers' nth-order endorsers could be topically distinct after several iterations. We here name the citing cascades that contain nth-order endorsers as *nth-order citing cascades*[3]; thus, the cascade proposed in this paper is typically a first-order cascade. It will be interesting to investigate the topic evolution of these publications, and some illuminating knowledge diffusion patterns through these paths might be identified. In addition to the idea from first- to nth-order cascades, another way to expand the definition of the citing cascade is to supplement certain constraints onto the cascade, such as content- and time-constrained citing cascades:

Content-constrained citing cascades: Based on the citing cascade defined in this paper (first-order cascade), we can filter the endorsers occurring in a cascade by using the topical relatedness between the owner and a given endorser. Specifically, future studies can choose a threshold and simply include the endorsers whose topical similarity with the owner is larger than the threshold and exclude those are not. By doing this, the research areas of the publications in a given cascade will be more focused and denser.

[3] Essentially the "citation cascade" mentioned in Min, Sun, and Ding (2017) and Min, Bu, Sun, and Ding (2018) should be categorized in this type.

Content-constrained citing cascades are particularly useful when we specialize in research about concept evolution and topic evolution.

Time-constrained citing cascades: In addition to filtering the endorsers by considering their content-level information, future researchers can also include the published time information of the owner and the endorsers. A potentially achievable way to do this is to reserve endorsers published within certain years after the owner has been published and exclude the others (Alvarez *et al.*, 2015).

Meanwhile, under the framework built by the current definitions, all of the citing publications are equally treated regardless of their differences, such as the published year of the citing publications and the topical relatedness between the citing and cited publications. Scientometricians should consider combining these attributes of publications with citing cascades. By doing so, the graph (network) built by the citing cascade will be more informative, in that it involves not only the nodes and edges in a graph but also the attributes of the nodes. Some regression models might also be adopted to solve related issues. From a purely scientometric perspective, some comparisons between scholarly relationships (e.g., co-citation [Bu, Ni, & Huang, 2017; Small, 1973; White & Griffith, 1981] and bibliographic coupling [Kessler, 1963; Zhao & Strotmann, 2008]) and citing cascades should be discussed.

Many other related issues could be studied productively in the near future. For example, an endorser might first act as a late endorser, but later also act as a connector. The dynamic change of the endorsers' roles could be interesting to understand the citing behavior-related issues, thus temporal analyses can be implemented by utilizing a framework that is similar to that of the current study. Moreover, since we present a basic description of citation depth in this paper, future work can consider citation depth more deeply and assess more thorough meanings of citing cascades. Finally, future researchers should focus more on how to understand bias in citing behavior by

modeling the details of knowledge diffusion through citing cascades and interpret "cascading behaviors" by utilizing both qualitative and quantitative approaches.

ACKNOWLEDGMENTS

This article is financially supported by the Major Program of Social Science Foundation in China (No. 17ZDA292). The authors acknowledge the Indiana University Pervasive Technology Institute for providing KARST, a high-performance computing system in Indiana University (Stewart *et al.*, 2017), that have contributed to the research results reported within this paper. This research was supported in part by the Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute, and in part by the Indiana METACyt Initiative. The Indiana METACyt Initiative at Indiana University was also supported in part by the Lilly Endowment, Inc. The authors would like to thank Pik-Mai Hui, Ludo Waltman, and Chao Lu for their insightful suggestions for improving this paper.

REFERENCES

Alvarez, R., Garcia, D., Moreno, Y., & Schweitzer, F. (2015). Sentiment cascades in the 15M movement. *EPJ Data Science, 4*(1), 6-18.

Anderson, L.R., & Holt, C.A. (1997). Information cascades in the laboratory. *The American Economic Review*, 847-862.

Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J., & Tiwari, M. (2015). Global diffusion via cascading invitations: Structure, growth, and homophily. In *Proceedings of the 24th international conference on World Wide Web* (pp. 66–76), May 18-22, 2015, Florence, Italy.

Bakshy, E., Hofman, J.M., Mason, W.A., & Watts, D.J. (2011). Everyone's an

influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 65-74), February 9-12, 2011, Hong Kong, China.

Baños, R.A., Borge-Holthoefer, J., & Moreno, Y. (2013). The role of hidden influentials in the diffusion of online information cascades. *EPJ Data Science*, *2*(1), 6-21.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Bianconi, G., & Barabási, A.-L. (2001). Competition and multiscaling in evolving networks. *Europhysics Letters, 54*(4), 436.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy, 100*(5), 992-1026.

Borge-Holthoefer, J., Baños, R.A., González-Bailón, S., & Moreno, Y. (2013). Cascading behavior in complex socio-technical networks. *Journal of Complex Networks*, *1*(1), 3–24.

Brancheau, J.C., & Wetherbe, J.C. (1990). The adoption of spreadsheet software: Testing innovation diffusion theory in the context of end-user computing. *Information Systems Research, 1*(2), 115-143.

Bu, Y., Ni, S., & Huang, W.-B. (2017). Combining multiple scholarly relationships with author cocitation analysis: A preliminary exploration on improving knowledge domain mappings. *Journal of Informetrics, 11*(3), 810-822.

Buldyrev, S.V., Parshani, R., Paul, G., Stanley, H.E., & Havlin, S. (2010). Catastrophic

cascade of failures in interdependent networks. *Nature*, 464, 1025-1028.

Cha, M., Benevenuto, F., Ahn, Y.-Y., & Gummadi, K.P. (2012). Delayed information cascades in Flickr: Measurement, analysis, and modeling. *Computer Network, 56*(3), 1066-1076.

Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., & Leskovec, J. (2014). Can cascades be predicted? In *Proceedings of the 23rd international conference on World Wide Web* (pp. 925-936), April 7-11, 2014, Seoul, Korea.

Clauset, A., Shalizi, C.R., & Newman, M.E.J. (2009). Power-law distributions in empirical data. *Society for Industrial and Applied Mathematics Review, 51*(4), 661-703.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*(368), 829-836.

Cui, P., Jin, S., Yu, L., Wang, F., Zhu, W., & Yang, S. (2013). Cascading outbreak prediction in networks: a data-driven approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 901-909), August 11-14, 2013, Chicago, Illinois, U.S.A.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics, 5*(1), 187-203.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7*(3), 583-592.

Ding, Y., Yan, E., Frazho, A. and Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology, 60*(11), 2229-2243.

Duan, W., Gu, B., & Whinston, A.B. (2005). Analysis of herding on the internet-an empirical investigation of online software download. In *Proceedings of the Eleventh Americas Conference on Information Systems* (pp. 488-492), August 11-15, 2005, Omaha, Nebraska, U.S.A.

Galstyan, A., & Cohen, P. (2007). Cascading dynamics in modular networks. *Physical Review E, 75(*3), 36109.

Golub, B., & Jackson, M.O. (2010). Using selection bias to explain the observed structure of internet diffusions. *Proceedings of the National Academy of Sciences of the United States of America, 107*(24), 10833-10836.

González-Bailón, S., Borge-Holthoefer, J., & Moreno, Y. (2013). Broadcasters and hidden influentials in online protest diffusion. *American Behavioral Scientist, 57*(7), 943-965.

Hisakado, M., & Mori, S. (2015). Information cascade, Kirman's ant colony model, and kinetic Ising model. *Physical A: Statistical Mechanics and Its Applications*, 417, 63-75.

Hisakado, M., & Mori, S. (2016). Information cascade on networks. *Physical A: Statistical Mechanics and Its Applications*, 450, 570-584.

Huang, Y., Bu, Y., Ding, Y., & Lu, W. (2018). Direct citations between citing publications. *arXiv*.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10-25.

Kleinberg, J. (2007). Cascading behavior in networks: Algorithmic and economic issues. *Algorithmic Game Theory*, *24*, 613-632.

Kostka, J., Oswald, Y. A., & Wattenhofer, R. (2008). Word of mouth: Rumor dissemination in social networks. In *Proceedings of the 2008 structural information and communication complexity* (pp. 185-196), June 17-20, 2008, Villars-sur-Ollon, Switzerland.

Kuhn, T., Perc, M., & Helbing, D. (2014). Inheritance patterns in citation networks reveal scientific memes. *Physical Review X, 4*(4), 041036.

Lai, Y.-C., Motter, A.E., & Nishikawa, T. (2004). Attacks and cascades in complex networks. *Lecture Notes in Physics*, 650, 299-310.

Leskovec, J., Adamic, L.A., & Huberman, B.A. (2007a). The dynamics of viral marketing. *ACM Transactions on the Web, 1*(1), 5.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., & Hurst, M. (2007b). Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM international conference on Data Mining* (Vols. 1-0, pp. 551-556), Minneapolis, Minnesota, U.S.A.

Leskovec, J., & Singh, A. (2005). Measuring cascading behavior in a recommendation network. Retrieved from http://www.cs.cmu.edu/~jure/pub/old/cascade.final.pdf

Leskovec, J., Singh, A., & Kleinberg, J. (2006). Patterns of influence in a recommendation network. In *Proceeding of the Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 380–389), April 9, 2006, Singapore City, Singapore.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., & Hurst, M. (2007). Patterns of Cascading Behavior in Large Blog Graphs. In *Proceedings of the 2007 SIAM international conference on Data Mining* (Vols. 1-0, pp. 551-556), Minneapolis, Minnesota, U.S.A.

Li, C., Ma, J., Guo, X., & Mei, Q. (2017). DeepCas: An end-to-end predictor of information cascades. In *Proceedings of the 26th international conference on World Wide Web* (pp. 577-586), April 3-7, 2017, Perth, Australia.

Liben-Nowell, D., & Kleinberg, J. (2008). Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences of the United States of America, 105*(12), 4633-4638.

Merton, R.K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159*(3810), 56-63.

Min, C., Bu, Y., Sun, J., & Ding, Y. (2018). Is scientific novelty reflected in citation patterns?. *Proceedings of the 81st Annual Meeting of the Association for Information Science and Technology, 55*(1).

Min, C., Sun, J., & Ding, Y. (2017). Quantifying the evolution of citation cascades. *Proceedings of the Association for Information Science and Technology, 54*(1), 761-763.

Newman, M.E.J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 46*(5), 323-351.

Perc, M. (2010). Zipf's law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenia's research as an example. *Journal of Informetrics, 4*(3), 358-364.

Perc, M. (2013). Self-organization of progress across the century of physics. *Scientific Reports*, 3, 1720.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the*

National Academy of Sciences of the United States of America, 105(45), 17268-17272.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.J.P., & Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th international conference on World Wide Web* (pp. 243-246), May 18-22, 2015, Florence, Italy.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265-269.

Stewart, C. A., Welch, V., Plale, B., Fox, G., Pierce, M., Sterling, T. (2017). Indiana University Pervasive Technology Institute. Bloomington, Indiana. https://doi.org/10.5967/K8G44NGB.

Sun, E., Rosenn, I., Marlow, C., & Lento, T.M. (2009). Gesundheit! Modeling contagion through Facebook news feed. In *Proceedings of the international AAAI conference on Weblogs and Social Media* (pp. 146-153), May 17-20, 2009, San Jose, California, U.S.A.

Trueman, B. (1994). Analyst forecasts and herding behavior. *The Review of Financial Studies, 7*(1), 97-124.

Walden, E., & Browne, G. (2002). Information cascades in the adoption of new technology. In *Proceedings of the twenty-third international conference on Information Systems* (pp. 435-443), December 15-18, 2002, Barcelona, Catalonia, Spain.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365-391.

Waltman, L., & Van Eck, N. J. (2015). Field-normalized citation impact indicators and

the choice of an appropriate counting method. *Journal of Informetrics, 9*(4), 872-894.

Waltman, L., & Yan, E. (2014). PageRank-related methods for analyzing citation networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice* (pp. 83-100). Springer.

Wang, C., Chen, W., & Wang, Y. (2012a). Scalable influence maximization for independent cascade model in large-scale social networks. *Data Mining and Knowledge Discovery, 25*(3), 545-576.

Wang, Z., Scaglione, A., & Thomas, R.J. (2012b). A Markov-transition model for cascading failures in power grids. In *Proceeding of the Forty-fifth Hawaii international conference on System Science* (pp. 2115-2124), January 8-12, 2012, Maui, Hawaii, U.S.A.

Watts, D.J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(9), 5766-5771.

White, H.D., & Griffith, B.C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*(3), 163-171.

Yu, B., & Fei, H. (2009). Modeling social cascade in the Flickr social network. In *Proceedings of the Sixth international conference on Fuzzy Systems and Knowledge Discovery* (vol. 7, pp. 566-570). August 14-16, 2009, Tianjin, China.

Zhao, D., Cappello, A, & Johnston, L. (2017). Functions of uni- and multi-citations: Implications for weighted citation analysis. *Journal of Data and Information Science, 2*(1), 51-69.

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in Information Science 1996-2005: Introducing author bibliographic coupling analysis. *Journal of the American Society for Information Science and Technology, 59*(13), 2070-2086.