Machine Learning (2020) 109:1205–1229
https://doi.org/10.1007/s10994-019-05828-9

Ranking by inspiration: a network science approach

Livio Bioglio[1] · Valentina Rho[1] · Ruggero G. Pensa[1]

Received: 24 June 2018 / Revised: 17 February 2019 / Accepted: 6 July 2019 / Published online: 21 August 2019

Abstract
Contagion processes have been widely studied in epidemiology and life science in general, but their implications are largely tangible in other research areas, such as in network science and computational social science. Contagion models, in particular, have proven helpful in the study of information diffusion, a very topical issue thanks to its applications to social media/network analysis, viral marketing campaigns, influence maximization and prediction. In bibliographic networks, for instance, an information diffusion process takes place when some authors, that publish papers in a given topic, influence some of their neighbors (coauthors, citing authors, collaborators) to publish papers in the same topic, and the latter influence their neighbors in their turn. This well-accepted definition, however, does not consider that influence in bibliographic networks is a complex phenomenon involving several scientific and cultural aspects. In fact, in scientific citation networks, influential topics are usually considered those ones that spread most rapidly in the network. Although this is generally a fact, this semantics does not consider that topics in bibliographic networks evolve continuously. In fact, knowledge, information and ideas are dynamic entities that acquire different meanings when passing from one person to another. Thus, in this paper, we propose a new definition of influence that captures the diffusion of inspiration within the network. We call it inspiration score, and show its effectiveness in detecting the most inspiring topics, authors, papers and venues in a citation network built upon two large bibliographic datasets. We show that the inspiration score can be used as an alternative or complementary bibliographic index in academic ranking applications.

Keywords Information diffusion · Bibliographic indexes · Citation networks · Topic modeling

Editors: Takuya Kida, Takeaki Uno, Tetsuji Kuboyama, Akihiro Yamamoto.

✉ Livio Bioglio
livio.bioglio@unito.it

Valentina Rho
valentina.rho@unito.it

Ruggero G. Pensa
ruggero.pensa@unito.it

[1] Department of Computer Science, University of Turin, Turin, Italy

1 Introduction

Contagion models and, in particular, stochastic epidemic models, have been widely studied in life sciences to understand and predict the spread of infectious diseases (Britton 2010). However, their implications are largely tangible in computational social science, machine learning and network science, due to their ability in explaining the dynamics of many social phenomena, such as the diffusion of ideas (Rogers 2003), the virality of certain posts or memes in social media (Yang and Zha 2013) and social influence (Cialdini and Trost 1998). Information diffusion, in particular, is a fundamental and very topical issue thanks to its applications to social media/network analysis (Bakshy et al. 2012), viral marketing campaigns (Leskovec et al. 2007), influence maximization (Chen et al. 2009) and prediction (Cui et al. 2011). An information diffusion process takes place when some active nodes (e.g., customers, social profiles, scientific authors) influence some of their inactive neighbors in the network and turn them into active nodes with a certain probability, and the newly activated nodes, in their turn, can progressively trigger some of their neighbors into becoming active (Gui et al. 2014). Information diffusion is similar to the spread of diseases in epidemiology and it has also been modeled as such (Daley and Kendall 1964) by considering influence as a contagion process. However the correct definition of "influence" strongly depends on the application. In mouth-to-mouth viral campaign, a user who buys a product at time T influences their neighbors if they buy the same product at time $T + \delta$. In social media, influence is the process that enables the diffusion of memes, (fake) news, viral posts across the network through different social actions such as likes, shares or retweets. In bibliographic networks, author a influences author b when a and b are connected by some relationship (e.g., collaboration, co-authorship, citation) and either b cites one of the papers published by author a, or author b publish in the same topic as author a (Gui et al. 2014). The latter definition, however, does not consider that influence in bibliographic networks is a complex phenomenon involving several scientific and cultural aspects. For instance, in scientific citation networks, the most cited papers are often seminal papers that introduce some topics (or some new aspects of a topic) for the first time. They are often cited "by default" and thus they spread in the network for very long periods. Moreover, in most existing works, influential topics are simply those ones that spread most rapidly in the network. Although this is generally a fact, this semantics does not consider that topics in bibliographic networks evolve continuously. In fact, knowledge, information and ideas are dynamic entities that acquire different meanings when passing from one person to another. For instance, "deep learning", a term invented in early 2000s, has known a rapid development and evolution that has influenced many research fields including semiconductor technology and circuits (Boguslawski et al. 2015; Coates et al. 2013; Seo and Seok 2015).

The above considerations also apply to bibliographic indexes, such as the very well known h-index by Hirsch (2005). Most indexes, in fact, are designed solely around the amount of citations received by papers of a given author (Lutz et al. 2008) or journal (da Silva and Memon 2017) to rate and rank them. Since scholars' career and success of editorial venues are heavily affected by these metrics, their societal and economic impact cannot be overlooked. In some cases, their misuse has even resulted in abnormal and questionable scientific behaviors.[1] Hence, capturing more realistic and effective influence patterns in citation networks may lead to a fairer measurement of authors and venues impact in science.

In this paper we address the problem of information diffusion in a bibliographic network by using the notion of *inspiration*. According to our definition, the most inspiring ideas are

[1] https://www.natureindex.com/news-blog/italian-scientists-increase-self-citations-in-response-to-promotion-policy.

those that evolve rapidly in the network by triggering fast citation rates. As an example, consider an author a_0 that publish a paper p_0 at initial time interval ΔT_0 of width δ. In the following time interval ΔT_1, the activated authors are those that publish a paper p_1 citing paper p_0. In the following time interval ΔT_2, the authors that publish a paper p_2 citing paper p_1 are activated. In general, we only consider citations from papers published at time interval ΔT_i to papers published at the previous time interval ΔT_{i-1}. The intuition that drives our setting is that, similarly to the most infectious diseases, the most inspiring ideas are those that spread from one author to another one by evolving and adapting to different settings and research fields. Our definition of inspiration can be easily and indifferently applied not only to papers, but also to authors, venues (such as conferences or journals) and topics. Therefore, for a given time interval width δ, our diffusion model enables the ranking of papers, authors, venues and topics according to an *inspiration score*: items that rank high for small values of δ are the most inspiring ones. Moreover, as regards the inspiration score of topics, differently from other state-of-the-art methods (Gui et al. 2014), we consider topics assigned to papers by an adaptive Latent Dirichlet Annotation (LDA) technique (Hoffman et al. 2010). According to this method, a paper p is said to cover a topic X if the LDA model states that p is generated by X with a probability greater than a threshold. By comparing our model to a standard diffusion model, we show the effectiveness of our framework on two large corpora: one consisting of about $155,000$ computer science papers and $225,000$ authors, and the other one involving $27,770$ physics papers co-authored by $11,002$ authors. In addition, we show that our inspiration score has a different and fairer semantics w.r.t. classic bibliographic indexes, since it does not depend on the amount of papers written or citations received, but, rather, it captures the impact and evolution of an idea within the research community.

The salient contributions of this paper, can be resumed as follows:

- We define *inspiration* as an alternative to influence in information diffusion;
- We introduce the definition of *inspiration score* as a measure of the inspiration speed: papers that trigger fast citation rates have a high inspiration score;
- We propose a topic analysis model enabling the ranking of topics according to their inspiration speed;
- We define a general framework for the fair assessment of the impact of individual papers, authors, venues and topics within the research community.

The remainder of the paper is organized as follows: related works are analyzed in Sect. 2; the inspiration diffusion model is presented in Sect. 3; Sect. 4 provides the report of our experiments; finally, we draw some conclusions and discuss some limitation of our approach in Sect. 5.

2 Related works

In this section, we briefly review the scientific literature by proposing the main differences between our proposal and other related research works in different domains.

Epidemiological models Information diffusion has been first regarded as a derivation of the process of disease propagation in contact networks (Hethcote 2000), a well-studied problem in epidemiology. In particular, the Susceptible-Infectious-Recovered (SIR) epidemic model (Keeling and Rohani 2008) is employed for modeling infectious diseases that confer lifelong (or long-term) immunity, such as measles, rubella or chickenpox. In this model a susceptible node can become infected, because of the presence of infectious nodes, and an infectious node can naturally recover after few time, gaining immunity to the disease. For a literature survey on the subject, the reader may refer to Britton (2010). The SIR model has been applied to information spreading since early years (Daley and Kendall 1964; Maki and Thompson 1973). In such models, susceptible individuals do not know the information item, then are susceptible to be informed; infectious individuals know and spread the information item, while recovered individuals already know the information item but do not spread it anymore. Sudbury (1985) finds that in a complete random network, i.e., a homogeneous network, a rumor can only spread to around the 80% of the total population; more recently Zanette (2002) calculated that such percentage is lower than 80% in small-world networks. Zhou et al. (2007) found that the number of nodes reached by the rumor depends on the topological structure of the network. Moreno et al. (2004) show that the density of susceptible nodes at the end of the process decays exponentially with the value of their degree. An extension of this model (Nekovee et al. 2008) also allows spontaneous recovery, justified as forgetting mechanism. In this case, the model behave more similarly to the classical SIR model.

Influence maximization An obvious application of information diffusion stands in the domain of marketing, where diffusion models are used to understand the process of information spread among potential customers with the goal of improving viral marketing campaigns (Goldenberg et al. 2001). Leskovec et al. (2007), for instance, mathematically characterize the propagation of products recommendation in the network of individuals. To this purpose, influence maximization aim at identifying the minimum set of influential individuals (called seed nodes) that can maximize the diffusion of information or behaviours through a social network (Aral and Dhillon 2018). The most recent contributions in this field try to define online methods for identifying the most influential nodes in dynamic networks (Wang et al. 2017), or to jointly finding the top seed nodes and the top relevant tags for targeted influence maximization in a social networks (Ke et al. 2018).

Topic diffusion Besides viral marketing studies, the success of Web 2.0 and online social networks has also boosted researches on topic diffusion. Gruhl et al. (2004a, b) leverage the theory of infectious diseases to capture the structure of topics and analyze their diffusion in the blogsphere. Yang and Counts (2010b) analyze Twitter by constructing a model that captures the speed, scale, and range of information diffusion. The same authors compare the diffusion patterns within Twitter and a weblog network, finding that Twitter's network is more decentralized and connected locally (Yang and Counts 2010a). Barbieri et al. (2013) define a novel and more accurate information propagation model: the authors propose a topic-aware extensions of the well-known Independent Cascade and Linear Threshold models (Kempe et al. 2003) by taking into account authoritativeness, influence and relevance. Topic evolution has also been regarded as extensions of the Latent Dirichlet Allocation (LDA) or the Probabilistic Latent Semantic Analysis algorithms (Gohr et al. 2009). He et al. (2009) leverage citations to address the problem of topic evolution analysis on scientific literature. When detecting topics in a collection of new papers at a given time instant, they also consider citations to previously published papers and propose a novel LDA-based topic modeling technique named Inheritance Topic Model. Kim et al. (2018) employ a citation influence topic model based on topical inheritance between cited and citing papers and analyze topic evolution in a circumscribed biological literature scenario. In our work, we adopt a similar solution, but we look at topic evolution from the information diffusion perspective, by computing a ranking of most inspiring topics, defined as those topics for which we observe a rapid evolution and inspiration in the network.

Citation networks Digital libraries and bibliographic networks have also taken advantage of information diffusion studies. Thanks to the availability of data sets of unprecedented size many studies have analyzed citation, co-authorship or co-participation networks to identify patterns of diffusion and influence, and to rank authors. Radicchi et al. (2009) define an author ranking method based on a diffusion algorithm that mimics the spreading of scientific credits on the network. Shi et al. (2009), instead, study the structural features of the information paths in the citation networks of publications in computer science. Among their findings, they discover that citing more recent papers corresponds to receiving more citations in turn. Gui et al. (2014) propose to model information diffusion in multi-relational bibliographic networks, by distinguishing different types of relationships. In addition, they propose a method to learn the parameters of their model leveraging real publication logs. Differently from all these works, we focus on diffusion and evolution of ideas by leveraging explicit citations in bibliographic networks.

Bibliographic metrics The availability of large (open) digitalized libraries and bibliographic data has also fostered the research on scientometrics, where one of the goal is to assess the impact of articles, authors, journals and institutes by leveraging citation patterns. In one of the most famous scientometric works, Hirsch (2005) proposes the h-index, one of the most used and most criticized bibliographic indexes. According to Hirsch, a scholar with an index of h has published h papers each of which has been cited in other papers at least h times. The predictive power of the h-index has also be demonstrated (Hirsch 2007), although its significance in ranking scientists has been questioned (Dorogovtsev and Mendes 2015), so that many other proposals try to address its limitations (Lutz et al. 2008), by leveraging centrality measures (Senanayake et al. 2014), among others. Although we do not address specifically the typical objectives of scientometric research, we show that a fair assessment of authors or editorial venues cannot be achieved without considering explicit citation patterns in the bibliographic networks.

3 Inspiration propagation

In this section we present the theoretical framework defining the notion of *inspiration* in a citation network. Initial inspiration takes place when a paper p written by an author a_0 is cited by another author a_1, for the first time. Successively, inspiration consists in another author a_2 citing at least one of the papers written by a_1 that cites, in its turn, the paper p written by a_0. Intuitively, an *inspiring paper* is a paper that is highly cited by other inspiring papers. Hence, the most inspiring papers are those that triggers fast inspiration mechanisms. Before providing the details of our framework, in the following we first introduce the mathematical background of our theoretical framework. Then we describe the diffusion model and define inspiration. Finally, we detail how to compute the inspiration score for each given author, venue or topic.

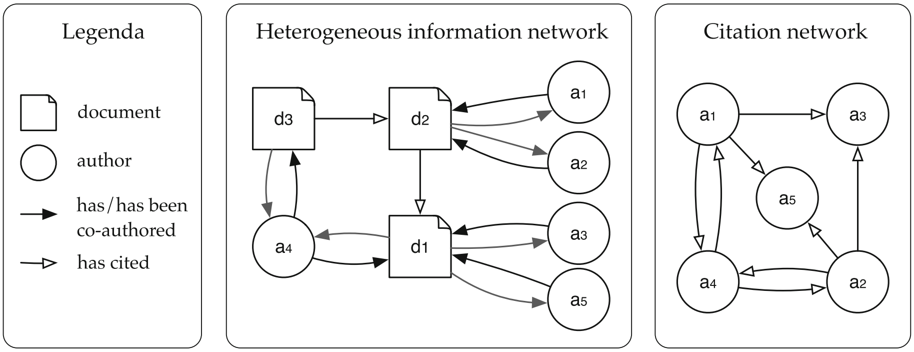


Fig. 1 A heterogenous information network and the corresponding citation network (Bioglio et al. 2017)

3.1 Information networks

We consider a set of n documents $D = d_1, \ldots, d_n$ and a set of K topics $Z = z_1, \ldots, z_K$. Each document $d_i \in D$ is characterized by a distribution of topics $\Theta_i =< \theta_{i1}, \ldots, \theta_{iK} >$, where $\forall i, k, 0 \leq \theta_{ik} \leq 1$ and $\sum_{k=1}^{K} \theta_{ik} = 1$ (how to compute the distribution Θ_i will be explained later on, in Sect. 3.4). Each document is authored by one or more authors belonging to the set $A = \{a_1, \ldots, a_N\}$ of all possible N authors. Moreover, each document d_i has a timestamp T_i corresponding to the publication date, and a venue b_i indicating the journal/conference/book where the paper has been published.

Authors and papers are part of a *heterogenous information network*, i.e., a directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = V^d \cup V^a$ and $\mathcal{E} = E^{ad} \cup E^{dd}$. Each $v_i^d \in V^d$ and $v_l^a \in V^a$ are, respectively, a vertex representing the i-th document $d_i \in D$ and a vertex representing the l-th author $a_l \in A$. Moreover, each $(v_l^a, v_i^d) \in E^{ad}$ is a directed edge meaning that author a_l has coauthored document d_i and each $(v_i^d, v_j^d) \in E^{dd}$ is a directed edge coding the fact that document d_i cites document d^j. Furthermore, E^{ad} is such that if $(v_l^a, v_i^d) \in E^{ad}$, then $(v_i^d, v_j^a) \in E^{ad}$ (i.e., each connection between documents and authors is reciprocal).

Within the heterogenous information network $\mathcal{G}(\mathcal{V}, \mathcal{E})$, we identify the *citation network* $G(V, E)$, where $V = V^a$ is the set of author vertices and $E = \{(v_h, v_l)\}$ is the set of directed citation edges. In particular, $(v_h, v_l) \in E$ iff there exists a path $path(v_h^a, v_l^a) = v_h^a \xrightarrow{ad} v_i^d \xrightarrow{dd} v_j^d \xrightarrow{ad} v_l^a$ within the information network $\mathcal{G}(\mathcal{V}, \mathcal{E})$. Roughly speaking, an edge (v_h, v_l) can be found in the citation network $\mathcal{G}(\mathcal{V}, \mathcal{E})$ iff author v_h has cited some (at least one) paper coauthored by v_l in one of the paper she coauthored. An example of heterogenous information network and its corresponding citation network is given in Fig. 1.

3.2 Diffusion model

Differently from most diffusion models that consider both co-authorship and citation links, our approach only considers explicit citations. In most existing approaches (such as the one presented by Gui et al. (2014)), the influence process takes place when an author publishes some paper on a given topic at time T and some of her neighbors publish any paper on the same topic at time $T + \delta$. Usually explicit citations are simply ignored, but they are crucial to understand the evolution and transformation of an idea across the network during a time period. Moreover, when explicit citations are ignored and heterogeneous links between authors are considered, the true semantics of propagation is less clear: influence may occur because of some external factors, e.g., the paper deals with a topic that is popular at publication time, the authors are part of the same consortium within a project, or they publish in the same topic just by chance. Instead, in our work, we propose to measure "inspiration" as an alternative to classic influence processes. Conversely speaking, inspiration takes place when an author cites another author explicitly in one of her papers, regardless of its topic. The general definition of inspiration is then as follows.

Definition 1 *(inspiration)* Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a heterogenous information network. Author $a_h \in A$ **is inspired by** author $a_l \in A$ $(a_l \neq a_h)$ iff there is a path $v_h^a \xrightarrow{ad} v_i^d \xrightarrow{dd} v_j^d \xrightarrow{ad} v_l^a$ in $\mathcal{G}$ s.t. $T_i \geq T_j$.

In the following we provide the theoretical details of our diffusion model. Let $\mathcal{T} = [T_0, T_n]$ be a time interval. We define a set $\Delta\mathcal{T} = \{\Delta T_0, \ldots, \Delta T_N\}$ of possibly overlapping time intervals over $\mathcal{T}$ s.t., $\forall t = 1 \ldots N$ $\Delta T_{t-1} \prec \Delta T_t$. We introduce the definitions of *initial inspiration* and *subsequent inspiration* as follows.

Definition 2 *(initial inspiration)* Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a heterogenous information network and $\Delta\mathcal{T} = \{\Delta T_0, \ldots, \Delta T_N\}$ a set of time intervals. Author $a_h \in A$ **is initially inspired by** author $a_l \in A$ $(a_l \neq a_h)$ iff there is a path $v_h^a \xrightarrow{ad} v_i^d \xrightarrow{dd} v_j^d \xrightarrow{ad} v_l^a$ in $\mathcal{G}$ s.t. $T_j \in \Delta T_0$ and $T_i \in \Delta T_1$.

According to this definition, the initial inspiration takes place when an author a_l publishes a document d_j during ΔT_0 $(T_j \in \Delta T_0)$, and another author a_h publishes a document d_i, during the following time interval ΔT_1 $(T_i \in \Delta T_1)$. Notice that we do not impose any constraint on the topic covered by document d_i. Let us now introduce the definition of *subsequent inspiration*.

Definition 3 *(subsequent inspiration)* Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a heterogenous information network and $\Delta\mathcal{T} = \{\Delta T_0, \ldots, \Delta T_N\}$ a set of time intervals. Author $a_h \in A$ **is subsequently inspired by** author $a_l \in A$ $(a_l \neq a_h)$ at time ΔT_t iff there is a path $v_h^a \xrightarrow{ad} v_i^d \xrightarrow{dd} v_j^d \xrightarrow{ad} v_l^a$ in $\mathcal{G}$ s.t. $T_j \in \Delta T_{t-1}$, $T_i \in \Delta T_t$ and a_l has been initially/subsequently inspired by another author $a_m \in A$ $(a_m \neq a_h$ and $a_m \neq a_l)$ during ΔT_{t-1}.

It can be noticed that this definition is recursive, meaning that the subsequent inspiration occurs when an author a_h has cited an author a_l that has been either subsequently inspired or initially inspired by a third author a_m in the previous time interval. Moreover, according to our diffusion model, inspiration takes place when a citation occurs between two consecutive time intervals. Even though this may appear a strong constraint, we recall that the definition of the set $\Delta\mathcal{T}$ of time interval is very general. In particular, we introduce two parameters $\delta > 0$ and $\gamma \geq 0$ $(\gamma < \delta)$, representing respectively the size of a sliding time window and the overlap between two consecutive time windows. Given these two parameters and a time interval $\mathcal{T} = [T_0, T_n]$, we define $\Delta\mathcal{T} = \{\Delta T_0, \ldots, \Delta T_N\}$ in such a way that $T_t = \big[T_0 + t(\delta - \gamma), T_0 + t(\delta - \gamma) + \delta\big)$, for $t = 0, \ldots, N$ with $N = \lceil \frac{T_n - (T_0 + \delta - 1)}{\delta - \gamma} \rceil$.

The rationale or the two parameters is as follows: given a time interval ΔT_i, the larger δ, the more papers are likely to cite papers of the previous time interval ΔT_{i-1}; the larger γ, the higher the probability that papers published during time interval ΔT_i will be cited by papers published during interval ΔT_{i+1}. Consequently, highly inspiring papers are those that trigger fast citation rates according to Definitions 2 and 3, i.e., when both δ and γ are small.

3.3 Computation of the inspiration score

We now describe how to assign a score value to each author, venue, topic or document depending on how fast inspiration spread across the network. To this purpose, for a given author a, venue b, topic z_k or document d, and a given set of time intervals $\Delta\mathcal{T} = \{\Delta T_0, \dots, \Delta T_N\}$ we measure the cumulative number of new authors inspired at each time interval, according to the definitions of inspiration given in Sect. 3.2. In the following, we first address the problem of identifying the initial seed of authors, which is slightly different, depending on which item the score is computed.

- *Initial seed for topic ranking* Given the heterogenous information network $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and a threshold $\tau \in [0, 1]$, we call $A(z_k)_0 = \{a_h \mid \exists(v_h^a, v_i^d) \in E^{ad} \wedge T_i \in \Delta T_0 \wedge \theta_{ik} > \tau\}$ the set of authors that publish a paper on topic z_k during ΔT_0.
- *Initial seed for venue ranking* Given the heterogenous information network $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and a venue b, we call $A(b)_0 = \{a_h \mid \exists(v_h^a, v_i^d) \in E^{ad} \wedge T_i \in \Delta T_0 \wedge b_i = b\}$ the set of authors that publish a paper d_i in venue $b_i = b$ during ΔT_0.
- *Initial seed for author ranking* Given the heterogenous information network $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and an author a, $A(a)_0 = \{a\}$. It is worth noticing that, in this case, the initial seed consists of the sole author a.
- *Initial seed for document ranking* Given the heterogenous information network $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and a document d, $A(d)_0 = \{a_h \mid \exists(v_h^a, v_i^d) \in E^{ad} \wedge T_i \in \Delta T_0\}$, that is the set of authors of document d.

Then, we define $A_1 = \{a_h \mid \exists a_l \in A(\star)_0 \ \ s.t. \ \ a_h \text{ is initially inspired by } a_l\}$ and, $\forall t = 2, \dots, N$, $A_t = \{a_h \mid \exists a_l \in A_{t-1} \ \ s.t. \ \ a_h \text{ is subsequently inspired by } a_l \text{ during } \Delta T_t\}$. In a nutshell, A_1 is the set of initially inspired authors, $A_2, \dots, A_N$ are the sets of subsequently inspired authors. In the definition of set A_1, $A(\star)_0 \in \{A(z_k)_0, A(b)_0, A(a)_0, A(d)_0\}$.

Finally, we construct a set of two-dimensional points $\{(t, y_t)\}$, $t = 1, \dots, N$ where $y_t = |A_t|$ for $t = 1$ and $y_t = |A_{t-1} \cup A_t|$ for $t = 2, \dots, N$. We use this set to compute a linear function $y = \hat{\sigma}t + \hat{c}$ by solving the following linear regression problem

$$(\hat{\sigma}, \hat{c}) = \arg\min_{\sigma, c} \sum_{t=1}^{N} (y_t - c - \sigma t)^2 \tag{1}$$

using the least squares method.

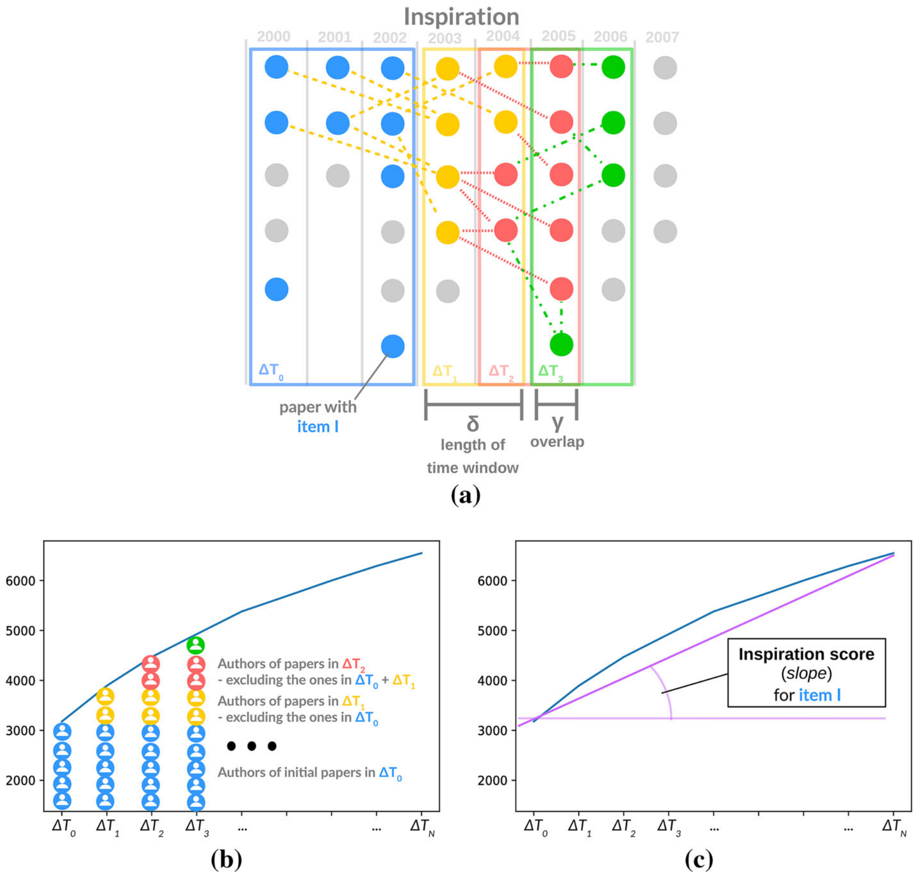
The *inspiration score* value is then defined as the slope $\hat{\sigma}$ of the linear function $y = \hat{\sigma}x + \hat{c}$ obtained by solving Equation 1. More formally:

Definition 4 *(topic inspiration score)* Given a heterogenous information network $\mathcal{G}(\mathcal{V}, \mathcal{E})$, a topic z_k, and a set of time intervals $\Delta\mathcal{T} = \{\Delta T_0, \dots, \Delta T_N\}$, the inspiration score of z_k, called $\mathrm{IR}(\mathcal{G}, \Delta\mathcal{T}, z_k)$, is given by

$$\mathrm{IR}(\mathcal{G}, \Delta\mathcal{T}, z_k) = \hat{\sigma} \tag{2}$$

where $\hat{\sigma}$ is the solution of the linear regression problem given in Equation 1 with $A_0 = A(z_k)_0$.

The *venue inspiration score*, noted $\mathrm{IR}(\mathcal{G}, \Delta\mathcal{T}, b)$, *author inspiration score*, noted IR $(\mathcal{G}, \Delta\mathcal{T}, a)$, and *document inspiration score*, noted $\mathrm{IR}(\mathcal{G}, \Delta\mathcal{T}, d)$, are computed by simply considering $A_0 = A(b)_0$ (resp. $A_0 = A(a)_0$ or $A_0 = A(d)_0$) as initial seeds. Notice that, by varying parameters δ and γ, which define the width and overlap of time intervals in $\Delta\mathcal{T}$, different values of inspiration score can be obtained. Moreover, in the author and venue


Inspiration
2000
2001
2002
2003
2004
2005
2006
2007
ΔT0
ΔT1
ΔT2
ΔT3
paper with item I
δ length of time window
γ overlap
6000
5000
4000
3000
2000
Authors of papers in ΔT2 - excluding the ones in ΔT0 + ΔT1
Authors of papers in ΔT1 - excluding the ones in ΔT0
Authors of initial papers in ΔT0
ΔT0 ΔT1 ΔT2 ΔT3 ... ... ΔTN
Inspiration score (slope) for item I

(a)

(b)

(c)

Fig. 2 Graphical representation of how our method works. **a** Select all the papers with item (venue, author or topic) I in ΔT_0, and let inspiration propagates among papers; **b** calculate the cumulative growth of the number of new authors (i.e. that are not authors of papers in previous time windows); **c** compute the inspiration score as slope of linear fitting of such cumulative

inspiration score, by also imposing a constraint on topics, it is possible to compute the score of an author or venue in relation to a specific topic

A graphical representation of our method for a generic item (venue, author or topic) is reported in Fig. 2.

3.3.1 Topic diffusion score

In order to compare our ranking method to the usual idea of topic diffusion, for each topic we also compute a *diffusion score* value as follows. For each time interval $\Delta T_t \in \Delta \mathcal{T}$ we set $A'_t = \{a_h \mid \exists (v^a_h, v^d_i) \in E^{ad} \wedge T_i \in \Delta T_t \wedge \theta_{ik} > \tau\}$, i.e., A'_t is the set of authors that have published a paper on topic z_k during time interval ΔT_t. Then, we construct a set of two-dimensional points $\{(t, y'_t)\}$, $t = 1, \ldots, N$ where $y'_t = |A'_t|$ for $t = 1$ and $y'_t = |A'_{t-1} \cup A'_t|$ for $t = 2, \ldots, N$. Again, we fit these values to a linear function $y' = \hat{\sigma} t + \hat{c}$ and set the *diffusion score* $\mathrm{DR}(\mathcal{G}, \Delta \mathcal{T}, z_k)$ equal to the slope $\hat{\sigma}$.

Algorithm 1: Topic inference on unseen documents in Online LDA.

```
Initialize $\psi_k = 1,\ \forall k = 1, \dots, K$.
repeat
    for $k = 1, \dots, K$ do
        Set $\phi_{wk} \propto \exp\{\mathbb{E}_q[\log \theta_k] + \mathbb{E}_q[\log \beta_w]\}\ \forall w = 1, \dots, N$
        Set $\psi_k = \alpha + \sum_w^N \phi_{wk} n_w$
until $\frac{1}{N} \sum_k$ |change in $\gamma_k| < \epsilon$;
return $\boldsymbol{\psi}$
```

3.4 Topic extraction

In this section, we introduce the topic modeling technique that we adopt to determine the distribution of topics for each document $d_i \in D$. Topic extraction is performed using Latent Dirichlet Allocation (LDA), a generative probabilistic model of a corpus, that aims to describe a set of observations, e.g. textual documents, using a set of unobserved latent elements, e.g. topics. LDA considers each document as a distribution over latent topics and each topic as a distribution over terms. Given α as prior knowledge about topics distribution, LDA assumes the following generative process for each document d of a corpus: (1) draw a distribution over topics $\theta_d \sim$ Dirichlet(α), (2) for each word i in d draw a topic z_{di} from θ_d and draw the word w_{di} from z_{di}.

For our purposes we use a slightly modified version of LDA, named *Online LDA* (Hoffman et al. 2010). In fact, traditional LDA implementations are based on either variational inference or collapsed Gibbs sampling; both methods require to process the entire corpus in order to compute the topic model, and it is not possible to query the model with previously unseen documents. In contrast, Online LDA replaces the previously used inference methods with the stochastic variational inference technique that allows *online* training, update of an existing model with new documents and query for unseen documents. Algorithm 1 shows the procedure to infer topics assignment on a new document; the document is represented by a vector of terms occurrences n of length N; K is the number of topics in the LDA model, α is the Dirichlet prior, β is the topic-term distribution matrix, ϕ and ψ are the variational parameters that represent respectively the topic assignments for each word and the topic proportion for the document.

4 Experiments

Here we introduce the corpus of scientific documents employed for our experiments, and we present the results of our ranking method on topics, venues, authors and single documents, comparing them to common ranking methods adopted in literature according to the kind of item under analysis. In the following, we start by describing the dataset used in our experiments, and how it has been constructed from raw data. Then we dedicate a single section to each ranked item: topics, venues, authors and single documents; for each of them we select the values for parameters σ and γ that best correlate with the ranking found with all the other pairs of values, employing it for the final score of each item. Such score is then compared with the results of a common ranking method used for the item. We also provide a preamble that describes some detail on the topic extraction and labeling tasks. The last

Table 1 Datasets statistics

	acm-v8	dblp-v8	Merged	Selected
No. of papers	2,381,674	3,272,990	1,373,202	154,947
No. of complete papers	1,668,246	3,241,890	1,143,443	154,947
No. of venues names	265,149	11,553	6959	153
No. of authors	1,508,051	1,752,440	903,771	225,559
No. of out-citations	8,650,089	8,466,858	6,513,765	1,321,905
No. of in-citations	–	–	5,365,753	1,000,657

section is dedicated on briefly repeating the entire workflow (except for topics) on a different secondary dataset, analyzing the differences with the main dataset.

4.1 Description of the main dataset

The dataset used in our experiments is a subset of the Computer Science paper citation network. This dataset is created by automatically merging two datasets originally extracted through ArnetMiner (Tang et al. 2008): the *DBLP* and *ACM* citation networks.[2] The merge procedure is necessary because both datasets lack some information: the *ACM* dataset contains many abstracts and citations between documents, but venues do not follow any naming convention and authors are ambiguous; In *DBLP*, venues and authors are clearly identified, but abstracts are missing and citations contain repetitions. Some statistics on the datasets are shown in Table 1. Papers are considered *complete* if all basic information are present, i.e. title, abstract (ACM only), year, venue and at least one outgoing or incoming citation. The *merged* dataset has been obtained by matching ACM and DBLP entries as follows: two papers match if both title and list of authors are the same. Then, abstracts and citations are extracted from ACM data; authors, title and venue are extracted from DBLP data. Finally, the *selected* dataset considers only papers published in the context of a set of manually preselected venues in the period from 2000 to 2014, covering the following research area: *artificial intelligence*, *machine learning*, *pattern recognition*, *data mining*, *information retrieval*, *database* and *information management*. The *selected* dataset is available online.[3]

4.2 Text processing, topic extraction and labeling

The input data given to the topic extraction algorithm is obtained as the result of a cleaning and vectorization process performed on the concatenation of paper title and abstract. In particular, the cleaning module ignores terms that appears only once in the dataset and in more than 80% of the documents. A domain dependent stop word list is also excluded from topic computation. First, documents are pre-processed with NLP techniques that perform tokenization, lemmatization, stop word removal and term frequency computation in order to prepare the corpus for the topic modeling algorithm. For performing this task, we adopt a scalable and robust topic modeling library (Řehůřek and Sojka 2010) that enables the extraction of an adaptive set of topics using an online learning version of Latent Dirichlet Allocation (Hoffman et al. 2010).

[2] https://aminer.org/citation.

[3] Dataset encoded in ArnetMiner V8 format, https://github.com/rupensa/tranet.

Table 2 Example of extracted topic description and associated labels

Topic description	Labels
0.091*network + 0.058*neural + 0.025*input + 0.021*learning + 0.021*adaptive + 0.020*neuron + 0.017*dynamic + 0.014*function + 0.014*output + 0.011*nonlinear + ...	Artificial Neural Network, Artificial Neuron, -

Topic modeling is performed on all papers published between 2000 and 2004 that appear within the *selected* dataset using Latent Dirichlet Allocation, searching for $K = 50$ topics. The extracted topic model is then used to assign a weighted list of topics to all papers published between 2005 and 2014. We perform LDA on a time interval preceding the one used for analysis, instead of the whole corpus, because in this way we focus on well-established topics rather than on emerging ones. However this choice does not limit our findings: in fact, many research topics investigated during the last ten years (including, e.g., *deep learning*) have been faced for the first time in the first half decade of the 21st century.

For improving the readability of our model, we introduce a simple topic labeling step that associates, to each topic z_k represented by a weighted list of words, up to three labels. The labels are computed as the first three results obtained by querying Wikipedia with the set of most representative words for z_k. We identify as *most representative* the 6 words having a weight greater than 0.01 or, if the first set is empty, the top 3 words. An example of labels extracted with this method is shown in Table 2.

4.3 Setup of experiments

In our experiments, we calculate the ranking of topics according to their *inspiration score* setting as initial time window ΔT_0 from 2000 to 2004, while the following time windows cover a time interval from 2005 to 2014. For topics only, the threshold τ has been set to 0.2. Algorithms and scripts are implemented in Python, and data are stored in a MongoDB[4] database server. The source code and the dataset are available online:[5] the whole analysis process can be driven within an interactive Jupyter notebook.[6] The experiments are performed on a server with two 3.30GHz Intel Xeon E5-2643 CPUs, 128GB RAM, running Linux.

4.4 Determining best parameters

Different values for the length of time window δ and number of years of overlap between subsequent time windows γ can lead to different ranking of items. In order to calculate a ranking that can be considered reliable and stable among different values, we explore the space of parameters, letting their values vary in ranges $1 \leq \delta \leq 6$ and $0 \leq \gamma \leq \delta$, and we consider as final ranking the one that is the most similar to all the other ones. For an item i (author, topic, venue or document), we calculate the ranking for each pair of values in our parameter space according to the method of the item described in Sect. 3, then for each pair of parametric values (δ, γ) we compute the average Spearman correlation with all the other

[4] https://www.mongodb.com/.

[5] https://github.com/bioglio/inspiration_score.

[6] https://jupyter.org/.

values for the parameters of the same item, obtaining an index of how much the ranking obtained from the current pair of parameters is similar to all the other ones. The Spearman's rank coefficient assesses monotonic relationships between two series of values. Given a set of n objects $X = x_i \dots x_n$ and two functions $f : X \rightarrow \mathbb{R}$ and $g : X \rightarrow \mathbb{R}$, the Spearman's coefficient is computed as:

$$\rho = 1 - 6 \cdot \frac{\sum_{i=1}^{n} \left(rank_f(x_i) - rank_g(x_i)\right)^2}{n(n^2-1)} \tag{3}$$

where $rank_f(x_i)$ and $rank_g(x_i)$ are the rank of object x_i in the two series of function values computed for X. It basically captures the correlation between the two rankings of the objects and ranges between -1 (for inversely correlated sets of values) and $+1$ (for the maximum positive correlation). The average Spearman's correlation for a pair of parametric values (δ, γ) on item i is computed as

$$\bar{\rho}_i(\delta,\gamma) = \frac{1}{N! - 1} \sum_{\delta'=1}^{N} \sum_{\substack{\gamma'=0 \\ (\delta',\gamma') \neq (\delta,\gamma)}}^{\delta'} \rho(rank_i(\delta,\gamma), rank_i(\delta',\gamma')) \tag{4}$$

where $N = 6$ is the maximum value of γ in our experiments, $\rho(r, r')$ is the Spearman correlation between two lists, and $rank_i(\delta, \gamma)$ is the result of our ranking method for the item i on the pairs of parameters (δ, γ). The pair of parameters $(\hat{\delta}, \hat{\gamma})$ chosen for the final ranking of the item i is the one that maximize this value:

$$(\hat{\delta}, \hat{\gamma}) = \underset{\delta,\gamma}{\arg\max}\ \bar{\rho}(\delta,\gamma) \tag{5}$$

Figure 3 shows the average Spearman correlation for each pair of parameters calculated for the four items: the optimal values result $(\hat{\delta}, \hat{\gamma}) = (5, 3)$ for authors, single documents and topics, while $(\hat{\delta}, \hat{\gamma}) = (4, 2)$ for venues.

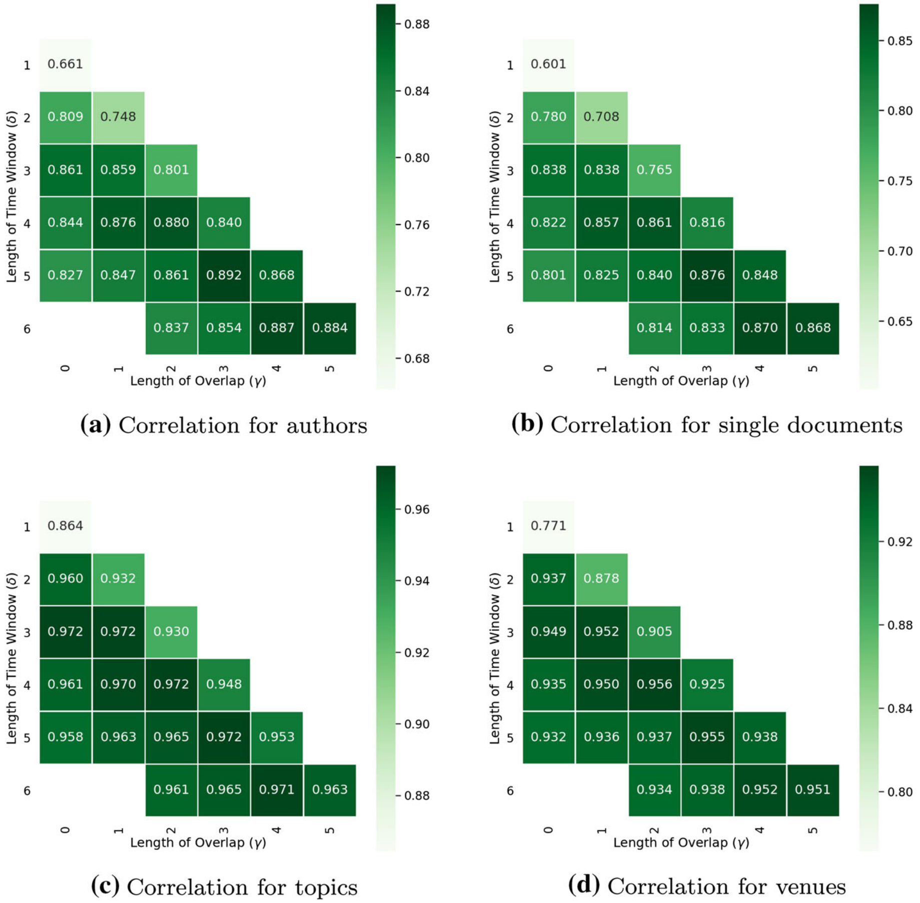
4.5 Ranking and comparison

In the present Section, for each item we interpret the results and we compare our ranking with the ones obtained by other methods commonly employed for ranking that particular item.

4.5.1 Authors

The number of authors with at least a publication in the initial time window contained in our dataset is 16, 245. The list of the top 10 authors according to our method and to the baseline of this item is reported in Table 3. According to the results only four scholars appear in both top-10s, underlying the difference between the two ranking measures. Interestingly, the authors that emerge from our ranking are known for their inspiring research in Latent Dirichlet Allocation (Michael I. Jordan and Andrew Y. Ng), data stream analysis (Pedro Domingos), collaborative filtering (George Karypis and Raymond J. Mooney) and support vector machines (Thorsten Joachims).

We compare the results calculated using our method with the ranking obtained according to the h-index (Hirsch 2005) of each author: such index has not been downloaded from external bibliographic sources, but it is calculated from the references present in our dataset, taking into account all the publications from 2000 to 2014. The Spearman's correlation between the



(a) Correlation for authors

(b) Correlation for single documents

(c) Correlation for topics

(d) Correlation for venues

Fig. 3 Average Spearman correlation calculated for each item (author, topic, venue or document) on each pair of parameters δ, on the y-axis, and γ, on the x-axis

ranking in our method and the one using h-index is 0.45: such moderately positive correlation indicates that h-index is only partially related to inspiration and the two indexes capture two distinct aspects of scientific productivity. To better investigate the reasons of this difference, we analyze to which extent these rankings are biased by other characteristics of authors. To this purpose, we also compare our ranking and the one obtained from h-index to the total number of papers published by the author, his seniority (calculated as the difference between years of his last and first publication), and three other widespread individual metrics: g-index, m-index and $i10$-index. $i10$-index counts the number of publications with at least 10 citations, m-index is the median number of citations received by papers in the Hirsch core (the set consisting of the first h articles, in order of decreasing citations), while, if we rank a set of articles in decreasing order of the number of citations received, the g-index (Egghe 2006) is the largest number such that the top g articles received altogether at least g^2 citations. As for h-index, all these values and metrics have been calculated using only the information in our dataset, and limited to time interval [2000, 2014]. Results in Fig. 4 show that our method has a low correlation with seniority and number of publications, and a medium correlation

Table 3 Top 10 ranking for authors, according to our method and h-index (calculated according to the information contained in our dataset)

Pos.	Our method		Baseline	
	Author	H-index	Author	H-index
1	Michael I. Jordan	16	Jiawei Han	30
2	Pedro Domingos	16	Christos Faloutsos	26
3	Wei-Ying Ma	25	ChengXiang Zhai	25
4	Jiawei Han	30	Wei-Ying Ma	25
5	Christos Faloutsos	26	Philip S. Yu	25
6	George Karypis	14	W. Bruce Croft	24
7	ChengXiang Zhai	25	Alon Y. Halevy	22
8	Thorsten Joachims	16	Jian Pei	22
9	Raymond J. Mooney	13	Qiang Yang	21
10	Andrew Y. Ng	12	Eamonn J. Keogh	20

In case of parity, authors with less publications come first

with g-index, while h-index is highly biased by them; although high productivity does not automatically turns into high citation rates, the high Spearman's correlation value, equal to 0.83, underlines the risk of using the sole h-index in ranking and assessing scholars. On the other hand, our method seems more correlated with m-index, that does not affect much the h-index, and i-10-index, at the same level of the h-index.

4.5.2 Venues

Among the 153 venues inserted in our dataset, we rank the 118 having at least one publication in the initial time window. The method used as a comparison is a modified version of Impact Factor. For a venue b, let $R(b, \Delta T_0, \Delta T') = \{d_i \mid \exists(v_j^d, v_i^d) \in E^{dd} \wedge T_j \in \Delta T_0 \wedge T_i \in \Delta T' \wedge b_j = b\}$ be the set of papers published during $\Delta T'$ that reference the papers published in venue b during ΔT_0, while $P(b, \Delta T_0) = \{d_i \mid d_i \in D \wedge T_i \in \Delta T_0 \wedge b_i = b\}$ is the set of papers published in venue b during ΔT_0: the Impact Factor for the venue b, $IF(b)$, is calculated as

$$IF(v) = \frac{R(b, \Delta T_0, \Delta T')}{P(b, \Delta T_0) \cdot |\Delta T'|} \tag{6}$$

where $|\Delta T'|$ is the number of years in the time interval $\Delta T'$; the result is an average Impact Factor for the years in $\Delta T'$, considering as source of references the years in ΔT_0. According to our experiments, we set $\Delta T_0 = [2000, 2004]$ and $\Delta T' = [2005, 2014]$.

The list of top 10 venues computed by our method and by the local baseline is shown in Table 4. Quite surprisingly, the most inspiring venues do not trigger the highest IF values. However, all the venues listed in the top part of our ranking are widely recognized conferences (KDD, CIKM, ICDM, SIGIR, WWW and IJCAI) or journals (Machine Learning, JMLR, TKDE, TPAMI). Instead, the venues ranked according to the Impact Factor, cover more general topics (e.g., expert systems, computer science surveys, networks, computational biology and bioinformatics) which receive many citations but are less inspiring in the research area considered in our study.

Table 4 Top 10 ranking for venues, according to our method and to the Impact Factor (calculated through our dataset)

Pos.	Venue	IF
Our method		
1	Int. Conf. on Knowledge Discovery and Data Mining (KDD)	4.64
2	Machine Learning	4.88
3	Journal of Machine Learning research	6.42
4	IEEE Trans. on Knowledge and Data Engineering	2.35
5	Conf. on Information and Knowledge Management (CIKM)	2.98
6	IEEE Int. Conf. on Data Mining (ICDM)	1.51
7	ACM Conf. on Research and Development in Inf. Ret. (SIGIR)	5.12
8	IEEE Transactions on Pattern Analysis and Machine Intelligence	3.86
9	The Web Conference (WWW)	7.03
10	Int. Joint Conf. on Artificial Intelligence (IJCAI)	1.66
Baseline		
1	Neurocomputing	162.70
2	Pattern recognition	67.20
3	Expert Systems with Applications	56.31
4	Knowledge-Based Systems	51.40
5	Conf. on Computer Vision and Pattern Recognition (CVPR)	51.33
6	Int. Journal of Geographical Information Science	11.90
7	Web Semantics	10.02
8	IEEE/ACM Trans. on Computational Biology and Bioinformatics	8.05
9	ACM Computing Surveys	7.18
10	IEEE Communications Letters	7.10

The Spearman's correlation between our ranking and the one that emerges from the modified version of Impact Factor is 0.51, showing a positive but moderate correlation between the two indexes. We also analyze the correlation between these rankings and two features of the venues, i.e., the number of publications in ΔT_0 and the number of citations received by the papers in ΔT_0 from papers in $\Delta T'$. Results in Fig. 4 show that, not surprisingly, both the rankings are highly correlated with the latter, while our ranking maintains a low positive correlation with the former (the custom Impact Factor, instead, exhibits a low negative correlation). Hence, although the two indexes show similar correlations with the number of publications and citations, they capture relatively different phenomena.

4.5.3 Topics

As described in Sect. 4.2, we perform LDA for extracting 50 different topics from the abstracts of the papers in our dataset: Table 5 reports the top 10 topics according to our method and to the baseline, i.e. the diffusion score presented in Sect. 3.3.1. We employ our diffusion score as baseline because the task of ranking topics is not common in the analysis of bibliographic networks, then in literature there is not an usual method available for such purpose: however, diffusion score is similar to the topic diffusion model adopted by Gui et al. (2014). We notice that the best 3 topics are the same for both methods, while the

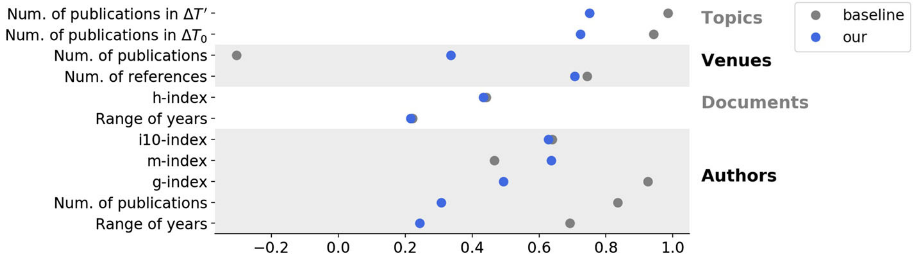


Fig. 4 Spearman correlation calculated on several features of items. Each feature is compared to the results from our method and from a baseline method, peculiar of the item under study: for Topics, number of citations received by documents published on topic in ΔT_0; for Venues, a modified version of Impact Factor; for Documents, number of references received; for Authors, h-index

Table 5 Top 10 ranking for topics, according to our method and diffusion ranking

Pos.	Our method Topic	Baseline Topic
1	19—Machine Learning	19—Machine Learning
2	42—Business Inf. Sys.	42—Business Inf. Sys.
3	36—Optimization	36—Optimization
4	43—Graph Databases	48—Image Processing
5	26—Information Retrieval	43—Graph Databases
6	41—Pattern Mining	10—Video Streaming
7	7—Time Complexity	15—Bioinformatics
8	6—Ontologies	14—Networking
9	12—Natural Language Processing	6—Ontologies
10	46—Statistical Relational Learning	26—Information Retrieval

remainder of the two rankings are sensibly different, i.e., some topics are missing in one of the two rankings, while shared topics are represented in different positions. Interestingly, topic *Information Retrieval* is ranked high for inspiration (in the $5th$ position), while it is ranked $10th$ according to diffusion. Our measure captures a real trend in Computer Science: the increasing research efforts in information retrieval have been driven by search engine and social media applications, as well as by Semantic Web technologies. Topic *Graph Databases* is also ranked higher by our technique. Research on this topic has been boosted by semantic database achievements in the last 15 years. Notice that our techniques also ranks *NLP* (Natural Language Processing) and *Pattern Mining* among the top 9 topics, coherently with the actual efforts in these domains pushed by the advances in sentiment analysis and other Semantic Web applications as well as in frequent itemset and sequence mining in the considered period: on the contrary, these topics are only ranked 13th and 19th according to standard diffusion metrics.

It is worth noting that, by analyzing the ranking of the top 10 topics based on the diffusion score, and their respective ranking based on our method in Table 5, we observe that some of the topics that have a relatively lower rank in the ranking-by-inspiration approach can be considered as application of Computer Science techniques. For instance, it is a fact that *Bioinformatics* (ranked seventh) has spread rapidly in the last 10 years. However, in our approach this topics gets a lower rank ($14th$): this can be explained by the fact that, in the

Table 6 Top 10 ranking for documents, according to our method and number of references; for each document, the main author and number of references are reported

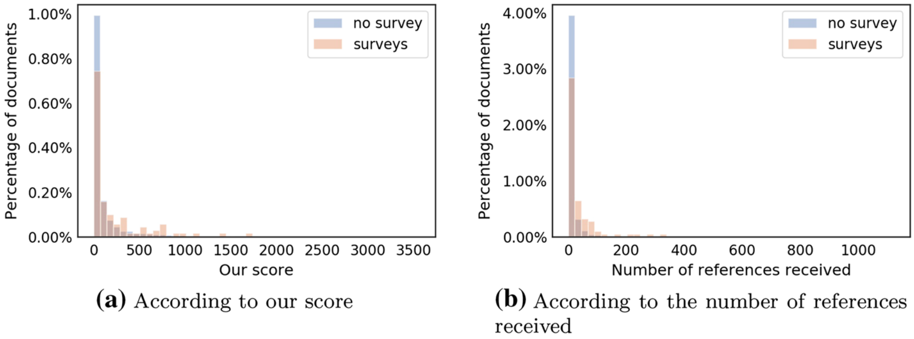
Pos.	Title	Authors	Refer.
Our method			
1	Optimizing search engines using clickthrough d...	Joachims T.	564
2	Latent Dirichlet Allocation	Blei D. M., ..	1,122
3	Text Classification from Labeled and Unlabeled...	Nigam K., ..	312
4	Machine learning in automated text categorizat...	Sebastiani F.	490
5	IR evaluation methods for retrieving highly re...	Järvelin K., ..	264
6	Web-scale information extraction in knowitall:...	Etzioni O., ..	102
7	RCV1: A New Benchmark Collection for Text...	Lewis D. D., ..	254
8	An Introduction to Variable and Feature Select...	Guyon I., ..	473
9	Document Language Models, Query Models, ...	Lafferty J. D., ..	209
10	Latent semantic models for collaborative filte...	Hofmann T.	175
Baseline			
1	Latent Dirichlet Allocation	Blei D. M., ..	1,122
2	Optimizing search engines using clickthrough d...	Joachims T.	564
3	Random Forests	Breiman L.	547
4	Machine learning in automated text categorizat...	Sebastiani F.	490
5	An Introduction to Variable and Feature Select...	Guyon I., ..	473
6	Content-Based Image Retrieval at the End of th...	Smeulders A., ..	446
7	Cumulated gain-based evaluation of IR techniques	Järvelin K., ..	441
8	Evaluating collaborative filtering recommender...	Herlocker J., ..	370
9	Laplacian Eigenmaps for Dimensionality Reducti...	Belkin M.	346
10	A survey of approaches to automatic schema mat...	Rahm E.	339

research areas under investigation, covering data mining and machine learning, papers in this multidisciplinary field are more likely to be inspired by (rather than to inspire) other research topics (such as, clustering, machine learning or pattern mining). The same observation applies to *Video Streaming* and *Image Processing*, in our method ranked $23th$ and $16th$, respectively.

Finally, as reported in Fig. 4, we found that both methods are highly correlated with the number of publications in topics during ΔT_0 and $\Delta T'$, although our method is less biased by these two parameters. Notice that the Spearman's correlation between the *inspiration score* and the *diffusion score* is 0.75.

4.5.4 Documents

Finally, we analyze the ranking of documents published during ΔT_0: the number of them having at least a citation in the following years, and then able to be ranked with our method, is $11,417$. The common manner for analyzing the success of a document is to count the number of citation it receives. Hence, Table 6 shows the titles of the best 10 documents found by our method and by this baseline. The most surprising result concerns the absence of the "Random Forests" paper in our ranking, while it is ranked third according to the number of citation received. This is probably due to the fact that, although Random Forests are widely used (and then cited), they have not inspired much further research in the scientific community



(a) According to our score

(b) According to the number of references received

Fig. 5 Distributions of documents recognized as surveys or reviews versus research papers

considered in the present study. We compare the correlation of these two methods with the age of document and the higher h-index among the authors of documents: we observe that, as summarized in Fig. 4, these methods show almost the same correlation with these features. Nonetheless, the Spearman's correlation between the two methods is equal to 0.65, showing that the information captured by the two rankings is not exactly the same.

As a last analysis on documents, we also analyze the performances of our score on surveys (or reviews) compared to research papers: we consider as "survey" all the publications whose titles contain one of the following keyphrases: *survey*, *review* or *state of the art*. Using such heuristic, we identify 181 papers. The distribution of our score, and the baseline score, on surveys versus research papers is reported in Fig. 5. According to the Kolmogorov-Smirnov test (Smirnov 1948), the two distributions are different (p-value < 0.01), for both our score and the number of references received, but in Fig. 5a we observe that our score is slightly less discriminant between the two types of articles, while Fig. 5b shows that, in general, a survey gains more references than a research paper.

4.6 Human assessment of the rankings

So far, we have showed how reasonable are the rankings computed using our method, according to some statistical measure and subjective interpretation of the results. Here, instead, we present the results of a survey conducted on the same community of researchers under study with the aim of measuring to what extent our ranking meets the experience (and expectations) of domain experts. We contacted directly 46 researchers working in the domains of machine learning, data mining, databases and information retrieval. More than half of them are experienced researchers (they have owned a Ph.D. for at least ten years), while less than one third can be considered as junior researchers (less than five years have passed since they have obtained their Ph.D.). The participants were asked to choose between two different ordered lists of top-25 authors (one obtained by using our inspiration score, one computed according to the h-index) and two different ordered list of top-20 venues (one resulting from our score, one resulting from a measure similar to the Impact Factor, as described in Sect. 4.5.2). The first question is formulated as follows: *"Which one of the following rankings of scientific authors best represents the most inspiring scholars, according to your opinion?"*. The second questions, instead, was asked as follows: *"Which one of the following rankings of scientific conferences and journals best represents the most inspiring scientific venues, according to your opinion?"*. To avoid any bias, the participants were not tell how the different ranking were computed. Moreover, we did not collect any personal information: hence the questionnaire, implemented using Google Forms,[7] is completely blind and anonymous.

Two weeks after we promoted the survey, we received 29 answers showing that, overall, our method outperforms the other standard measures both for authors and for venues. In detail, 55.2% of the respondents consider the inspiration score better than the h-index in ranking authors. As regards the venues, the results are even sharper: our ranking has been preferred by 96.6% of the participants. The results suggest that our score can be effectively proposed as an alternative measure to rank conference and journals, since the ranking it computes meets the perception of a great majority of the researchers involved in our study. Moreover, although the results are not as clear as for the venues, it can also be used to rank authors in association with other metrics. In particular, since our score has little correlation with productivity-oriented metrics (see Sect. 4.5.1), it can be used in combination with them to provide a multi-objective or multi-aspect performance comparison. Nonetheless, this study also implicitly suggests that an agreement on a universally valid measure for ranking scientific authors is far from being reached.

4.7 Analysis on the secondary dataset

In this section we present the results of the same experimentation on a different dataset, i.e. the citation graph of the e-prints uploaded on arXiv[8] in the field of high energy physics theory from January 1993 to April 2003:[9] it is composed by $27,770$ papers and $352,807$ references, authored by $11,002$ researchers. Differently from the main dataset, here we analyze the performances of our score only on authors, venues and documents: the exclusion of topics is due to the lack of text data for repeating the topic extraction procedure described in Sect. 4.2.

4.7.1 Setup of experiments

For the present dataset, we calculate the inspiration score setting as initial time window ΔT_0 from 1992 to 1994, while the following time windows cover a time interval from 1995 to 2003: more formally, $\Delta T_0 = [1992, 1994]$ and $\Delta T' = [1995, 2003]$. We also repeat the method described in Sect. 4.4 for determining the best values of the parameters of our inspiring score (the length of time window δ and the number of years of overlap between subsequent time windows γ): the optimal values result $(\hat{\delta}, \hat{\gamma}) = (5, 3)$ for venues, while $(\hat{\delta}, \hat{\gamma}) = (4, 2)$ for authors and single documents.

4.7.2 Authors

The number of authors under analysis, i.e. that have published at least a paper in the initial time window, is $3,672$: the lists of top 10 authors according to our score and to baseline (the h-index calculated with the references available in the dataset) are reported in Table 7. The Spearman's correlation between the two ranks is 0.49, very close to the one found in the main dataset; we also compare both rankings with the number of papers published by the author, his seniority, and the metrics g-index, m-index and $i10$-index, obtaining the correlations

[7] https://www.google.com/forms/about/.

[8] https://arxiv.org/.

[9] The dataset is made available within the SNAP project at https://snap.stanford.edu/data/cit-HepTh.html.

Table 7 Top 10 ranking for authors, according to our method and h-index (calculated according to the information contained in the dataset)

Pos.	Our method		Baseline	
	Author	H-index	Author	H-index
1	C. Vafa	54	E. Witten	63
2	J. Polchinski	34	C. Vafa	54
3	A.Sen	47	N. Seiberg	49
4	P. K. Townsend	35	N. A.Sen	47
5	M. J. Duff	28	A. A. Tseytlin	42
6	E. Witten	63	A. Strominger	39
7	C. M. Hull	21	I. R. Klebanov	38
8	J. X. Lu	13	M. R. Douglas	36
9	N. Seiberg	49	L. Susskind	36
10	R. R. Khuri	15	R. Kallosh	36

In case of parity, authors with less publications come first

Table 8 Top 10 ranking for venues, according to our method and to the Impact Factor (calculated through our dataset)

Pos.	Venue	IF
Our method		
1	Journal of Mathematical Physics	2.49
2	Physics Reports	50.17
3	International Journal of Modern Physics	2.86
4	Physical Review	8.31
5	Modern Physics Letters	1.73
6	Physical Review Letters	11.16
7	Communications in Mathematical Physics	3.36
8	Classical and Quantum Gravity	5.82
9	Journal of Geometry and Physics	4.07
10	Physics Letters	6.81
Baseline		
1	Physics Reports	50.17
2	Lecture Notes in Physics	27.78
3	Nuclear Physics B - Proceedings Supplements	20.58
4	Fortschritte der Physik	17.71
5	Nuclear Physics	14.03
6	Physical Review Letters	11.16
7	Chaos, Solitons & Fractals	8.67
8	Physical Review	8.31
9	Physics Letters	6.81
10	Classical and Quantum Gravity	5.82

in Fig. 6: interestingly, these correlations are quite similar to the ones found for the main dataset, with the only exception of the one between h-index and $i10$-index. In addition, our analysis shows a very high correlation between h-index and g-index. By looking at the names, our ranking correctly identify some of the pioneers in the second superstring revolution and

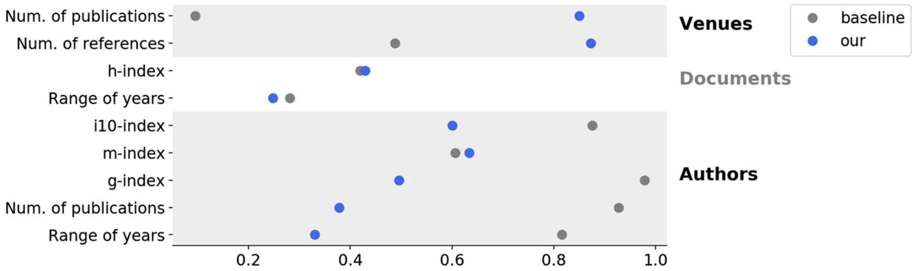


Fig. 6 Spearman correlation calculated on several features of items. Each feature is compared to the results from our method and from a baseline method, peculiar of the item under study: for Venues, a modified version of Impact Factor; for Documents, number of references received; for Authors, h-index

string solitons, and, in general, authors that have inspired some of the most important theories developed during the years under study.

4.7.3 Venues

Before analyzing the ranking of venues, we reduce the noise in data by considering only the 50 venues with at least 10 publications: among them, only 36 have at least one publication in the initial time window. The list of the top 10 venues according to our method and to the baseline, i.e. the Impact Factor calculated as in Equation 6, is reported in Table 8: again, our method does not reward journals with higher Impact Factor, as also showed by the low Spearman's correlation between these two rankings (0.3). Figure 6 shows the correlation of these rankings with the number of publications in ΔT_0 and the number of citations received by the papers in ΔT_0 from papers in $\Delta T'$: the correlation with our score is quite similar to the one found for the main dataset, while the Impact Factor exhibits a very different behavior, with a medium correlation with the number of references and almost no correlation with the number of publications. However, this could be a consequence of the fact that not all articles published in the considered venues have been uploaded on arXiv.

4.7.4 Documents

In this section we calculate the rank of the $5,802$ papers published during the initial time window, and we compare our score with the number of citations a paper has received: the list of top 10 publications for both methods is shown in Table 9. The Spearman's correlation between the two methods is 0.6, very close to the one calculated for the main dataset. The same happens for the correlation of these two methods with the age of document and the higher h-index among the authors of documents: Fig. 6 shows that we obtain values that are very close to the ones calculated for the main dataset. A closer look at the papers confirms that our method is able to track works that are at the basis of important research branches (e.g., the "N=2" superstring theory).

5 Conclusions

We have proposed a new definition of influence that takes into account the diffusion of inspiration within a citation network. We have defined a new influence measure, called

Table 9 Top 10 ranking for documents, according to our method and number of references; for each document, the main author and number of references are reported

Pos.	Title	Authors	Refer.
Our method			
1	A Strong Coupling Test of S-Duality	Vafa C., ..	290
2	Unity of Superstring Dualities	Hull C. M., ..	748
3	Dyon - Monopole Bound States, Self-Dual Harmon...	Sen A.	240
4	Monopole Condensation, And Confinement In N=2 ...	Seiberg N., ..	1,299
5	String Solitons	Duff M. J., ..	426
6	Combinatorics of Boundaries in String Theory	Polchinski J.	136
7	The World as a Hologram	Susskind L.	427
8	Strong-Weak Coupling Duality in Four Dimension...	Sen A., ..	282
9	Vacuum interpolation in supergravity via super...	Gibbons G. W., ..	172
10	Target Space Duality in String Theory	Giveon A.	421
Baseline			
1	Monopole Condensation, And Confinement In N=2 ...	Seiberg N., ..	1,299
2	Monopoles, Duality and Chiral Symmetry Breakin...	Seiberg N.	1,006
3	Unity of Superstring Dualities	Hull C. M., ..	748
4	The World as a Hologram	Susskind L.	427
5	String Solitons	Duff M. J., ..	426
6	Target Space Duality in String Theory	Giveon A.	421
7	Electric-Magnetic Duality in Supersymmetric No...	Seiberg N.	411
8	The Black Hole in Three Dimensional Space Time	Bañados M., ..	380
9	Phases of N=2 Theories In Two Dimensions	Witten E.	344
10	Simple Singularities and N=2 Supersymmetric Ya...	Klemm A.	314

inspiration score, that captures the inspiration speed of topics (extracted using an adaptive LDA technique), authors, venues and papers, within a given time interval. The *inspiration score* allows the discovery of the most inspiring items according to different levels of speed. We have shown experimentally the effectiveness of our measure in ranking authors, venues, papers and topics in a citation network built upon two large bibliographic datasets. Although the core application is the analysis of inspiration diffusion in citation networks, our methods can be also applied on other information networks, including patent and news, provided that a link between two documents can be inferred directly or indirectly.

Nonetheless, our approach has some limitations. First of all, measuring inspiration requires a rather long observation period, although setting parameters accurately, it could be still used to rank items in a short interval of time (e.g., four or five years). Second, as most bibliographic indexes based on citations, our approach does not take into account the reason why a paper cites another paper, which would require more accurate natural language understanding and learning techniques. Finally, although our measure may enable the computation of fairer performance indicators than standard measures such as h-index and Impact Factor, it cannot be computed trivially by just counting citations. However, as future work, we plan to expose our methodology to the public by deploying a dedicated web application.

Acknowledgements We wish to thank the anonymous reviewers for their valuable comments and suggestions. We also thank Prof. Igor Pesando for his support in interpreting the results on the high energy physics dataset. Finally, we thank all the researchers who participated in our survey for the assessment of part of the results. This work has been partially funded by Project MIMOSA (MultIModal Ontology-driven query system for the heterogeneous data of a SmArtcity, "Progetto di Ateneo Torino_call2014_L2_157", 2015-17).

References

Aral, S., & Dhillon, P. S. (2018). Social influence maximization under empirical influence models. *Nature Human Behaviour*, *2*, 375.

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. A. (2012). The role of social networks in information diffusion. In *Proceedings of WWW 2012* (pp. 519–528). ACM.

Barbieri, N., Bonchi, F., & Manco, G. (2013). Topic-aware social influence propagation models. *Knowledge and Information Systems*, *37*(3), 555–584.

Bioglio, L., Rho, V., & Pensa, R. G. (2017). Measuring the inspiration rate of topics in bibliographic networks. In *Discovery science—20th international conference, DS 2017*, Kyoto, Japan, October 15–17, 2017, *Proceedings. Lecture notes in computer science* (Vol. 10558, pp. 309–323). Springer.

Boguslawski, B., Sarhan, H., Heitzmann, F., Seguin, F., Thuries, S., Billoint, O., et al. (2015). Compact interconnect approach for networks of neural cliques using 3d technology. *Proceedings of IFIP/IEEE VLSI-SoC*, *2015*, 116–121.

Britton, T. (2010). Stochastic epidemic models: A survey. *Mathematical Biosciences*, *225*(1), 24–35.

Chen, W., Wang, Y., & Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 199–208). Paris, France, ACM, June 28–July 1.

Cialdini, R. B., & Trost, M. R. (1998). *Social influence: Social norms, conformity, and compliance* (Vol. 2, pp. 151–192). New York: McGraw-Hill.

Coates, A., Huval, B., Wang, T., Wu, D. J., Catanzaro, B., & Ng, A. Y. (2013). Deep learning with COTS HPC systems. In *Proceedings of ICML 2013, JMLR.org* (pp. 1337–1345).

Cui, P., Wang, F., Liu, S., Ou, M., Yang, S., & Sun, L. (2011). Who should share what?: Item-level social influence prediction for users and posts ranking. In *Proceeding of ACM SIGIR 2011* (pp. 185–194). ACM.

Daley, D. J., & Kendall, D. G. (1964). Epidemics and rumours. *Nature*, *208*, 1118.

Dorogovtsev, S. N., & Mendes, J. F. F. (2015). Ranking scientists. *Nature Physics*, *11*, 882–883.

Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics*, *69*(1), 131–152.

Gohr, A., Hinneburg, A., Schult, R., & Spiliopoulou, M. (2009). Topic evolution in a stream of documents. In *Proceedings of SIAM SDM 2009* (pp. 859–870). SIAM.

Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, *12*(3), 211–223.

Gruhl, D., Guha, R. V., Liben-Nowell, D., & Tomkins, A. (2004a). Information diffusion through blogspace. In *Proceedings of WWW 2004* (pp. 491–501). ACM.

Gruhl, D., Liben-Nowell, D., Guha, R. V., & Tomkins, A. (2004b). Information diffusion through blogspace. *SIGKDD Explorations*, *6*(2), 43–52.

Gui, H., Sun, Y., Han, J., & Brova, G. (2014). Modeling topic diffusion in multi-relational bibliographic information networks. In *Proceedings of CIKM 2014* (pp. 649–658). ACM.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, C. L. (2009). Detecting topic evolution in scientific literature: How can citations help? In *Proceedings of ACM CIKM 2009* (pp. 957–966). ACM.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review*, *42*(4), 599–653.

Hirsch, J. E. (2005). An index to quantify an individual?s scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, *104*(49), 19193–19198.

Hoffman, M. D., Blei, D. M., & Bach, F. R. (2010). Online learning for latent dirichlet allocation. *Proceedings of NIPS*, *2010*, 856–864.

Ke, X., Khan, A., & Cong, G. (2018). Finding seeds and relevant tags jointly: For targeted influence maximization in social networks. In *Proceedings of the 2018 international conference on management of data, SIGMOD conference 2018* (pp. 1097–1111). Houston, TX, USA, June 10–15, ACM 2018.

Keeling, M. J., & Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton: Princeton University Press.

Kempe, D., Kleinberg, J. M., & Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of ACM SIGKDD 2003* (pp. 137–146). ACM.

Kim, M., Baek, I., & Song, M. (2018). Topic diffusion analysis of a weighted citation network in biomedical literature. *JASIST*, *69*(2), 329–342.

Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007). The dynamics of viral marketing. *TWEB*, *1*(1), 5.

Lutz, B., Rüdiger, M., & Hans-Dieter, D. (2008). Are there better indices for evaluation purposes than the h index? a comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, *59*(5), 830–837.

Maki, D. P., & Thompson, M. (1973). *Mathematical models and applications: with emphasis on the social, life, and management sciences*. Upper Saddle River: Prentice-Hall.

Moreno, Y., Nekovee, M., & Pacheco, A. F. (2004). Dynamics of rumor spreading in complex networks. *Physical Review E*, *69*(6), 066130.

Nekovee, M., Moreno, Y., Bianconi, G., & Marsili, M. (2008). Theory of rumour spreading in complex social networks. CoRR abs/0807.1458.

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, *80*, 056103.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45–50).

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.

Senanayake, U., Piraveenan, M., & Zomaya, A. Y. (2014). The p-index: Ranking scientists using network dynamics. In *Proceedings of the international conference on computational science, ICCS 2014*. Cairns, Queensland, Australia, 10–12 June, 2014, Procedia computer science (Vol. 29, pp. 465–477). Elsevier.

Seo, J., & Seok, M. (2015). Digital CMOS neuromorphic processor design featuring unsupervised online learning. In *Proceedings of IFIP/IEEE VLSI-SoC 2015* (pp. 49–51). IEEE.

Shi, X., Tseng, B. L., & Adamic, L. A. (2009). Information diffusion in computer science citation networks. In *Proceedings of ICWSM 2009*. The AAAI Press

da Silva, J. A. T., & Memon, A. R. (2017). Citescore: A cite for sore eyes, or a valuable, transparent metric? *Scientometrics*, *111*(1), 553–556. https://doi.org/10.1007/s11192-017-2250-0.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, *19*(2), 279–281.

Sudbury, A. (1985). The proportion of the population never hearing a rumour. *Journal of Applied Probability*, *22*, 443–446.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. *Proceedings of KDD*, *2008*, 990–998.

Wang, Y., Fan, Q., Li, Y., & Tan, K. (2017). Real-time influence maximization on dynamic social streams. *PVLDB*, *10*(7), 805–816.

Yang, J., & Counts, S. (2010a). Comparing information diffusion structure in weblogs and microblogs. In *Proceedings of ICWSM 2010*. The AAAI Press.

Yang, J., & Counts, S. (2010b). Predicting the speed, scale, and range of information diffusion in twitter. In *Proceedings of ICWSM 2010*. The AAAI Press.

Yang, S., & Zha, H. (2013). Mixture of mutually exciting processes for viral diffusion. In *Proceedings of the 30th international conference on machine learning, ICML 2013*. (Vol. 28, pp. 1–9). Atlanta, GA, USA, 16-21 June 2013, JMLR.org, JMLR workshop and conference proceedings.

Zanette, D. H. (2002). Dynamics of rumor propagation on small-world networks. *Physical Review E*, *65*(4), 041908.

Zhou, J., Liu, Z., & Li, B. (2007). Influence of network structure on rumor propagation. *Physics Letters A*, *368*(6), 458–463.