Measuring the Inspiration Rate of Topics in Bibliographic Networks

Livio Bioglio, Valentina Rho, and Ruggero G. Pensa

Dept. of Computer Science, University of Turin, Turin, Italy
{livio.bioglio,valentina.rho,ruggero.pensa}@unito.it

Abstract. Information diffusion is a widely-studied topic 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 propose a measure of the inspiration rate called inspiration rank. Finally, we show the effectiveness of our measure in detecting the most inspiring topics in a citation network built upon a large bibliographic dataset.

Keywords: information diffusion · topic modeling · citation networks

1 Introduction

Information diffusion is a fundamental and widely-studied topic in many research fields, including computational social science, machine learning and network analytics, thanks to its applications to social media/network analysis [1], viral marketing campaigns [17], influence maximization [4] and prediction [6]. 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 [12]. Information diffusion is similar to the spread of diseases in epidemiology and it has also been modeled as such [7] 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 publishes in the same topic as author a [12]. 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 [3, 5, 20].

In this paper we address the problem of information diffusion in a bibliographic network by using the notion of *inspiring topics*. According to our definition, the most inspiring topics are 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 covering a given topic X at initial time interval t_0 of width δ. In the following time interval t_1, the activated authors are those that publish a paper p_1 citing paper p_0. In the following time interval t_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 t_i to papers published at the previous time interval t_{i-1}. Moreover, differently from other state-of-the-art methods, we consider topics assigned to papers by an adaptive Latent Dirichlet Annotation (LDA) technique [15]. 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. Therefore, for a given time interval width δ, our topic diffusion model enables the ranking of topics according to their inspiration rate: topics that rank high for small values of δ are the most inspiring ones.

The salient contributions of this paper can be summarized as follows: (1) we define *inspiration* as an alternative to influence in information diffusion; (2) we introduce the definition of *inspiration rank* as a measure of the topic inspiration rate: topics that trigger fast citation rates have a high inspiration rank; (3) we use an adaptive LDA technique for assigning topics to each paper; (4) we propose a topic analysis model enabling the ranking of topics according to their inspiration rate. By comparing our model to a standard diffusion model, we show the effectiveness of our framework on a large corpus consisting of about 155,000 scientific papers and 225,000 authors.

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

2 Related works

Information diffusion has been first regarded as a derivation of the process of disease propagation in contact networks [14], a well-studied problem in epidemiology. An obvious application 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 [9]. In [17], the authors mathematically characterize the propagation of products recommendation in the network of individuals.

Besides viral marketing studies, the success of Web 2.0 and online social networks has also boosted researches on topic diffusion. In [10, 11] the authors leverage the theory of infectious diseases to capture the structure of topics and analyze their diffusion in the blogsphere. In [25], Yang and Counts analyze Twitter by constructing a model that captures the speed, scale, and range of information diffusion. In [24], the same authors compare the diffusion patterns within Twitter and a weblog network, finding that Twitter's network is more decentralized and connected locally. In [2], a novel and more accurate information propagation model is defined: the authors propose a topic-aware extensions of the well-known Independent Cascade and Linear Threshold models [16] by taking into account authoritativeness, influence and relevance.

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. In [18], Radicchi *et al.* define an author ranking method based on a diffusion algorithm that mimics the spreading of scientific credits on the network. Shi *et al*, instead, study the structural features of the information paths in the citation networks of publications in computer science [21]. Among their findings, they discover that citing more recent papers corresponds to receiving more citations in turn. In [12], the authors 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 topic diffusion and evolution by leveraging explicit citations in bibliographic networks. Topic evolution has already been regarded as extensions of the Latent Dirichlet Allocation (LDA) or the Probabilistic Latent Semantic Analysis algorithms [8]. In [13] the authors 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

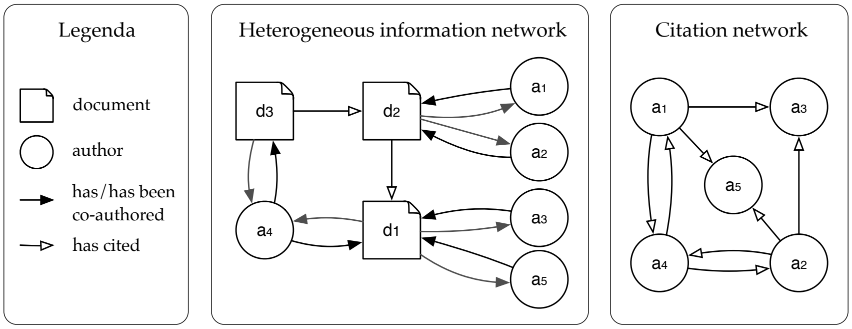


Fig. 1: A heterogenous information network and the corresponding citation network.

instant, they also consider citations to previously published papers and propose a novel LDA-based topic modeling technique named Inheritance Topic Model. 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 rate in the network.

3 Inspiration propagation

In this section we introduce the mathematical background and the theoretical framework of our ranking method.

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 $\sum_{k=1}^{K} \theta_{ik} = 1$. 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 ts_i corresponding to the publication date.

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 $G(V,E)$ iff author v_h has cited some (at least one) paper coauthored by v_l in one of the papers she coauthored. An example of heterogenous information network and its corresponding citation network is given in Figure 1.

In the following sections we describe the topic diffusion model adopted in our framework, as well as the topic modeling method used to associate topics with documents.

3.1 Topic diffusion model

Differently from most topic 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 in [12]), 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 a topic 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 topic 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. $ts_i \geq ts_j$.*

In the following we provide the theoretical details of our topic 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 topic-based inspiration* and *subsequent topic-based inspiration* for a given topic z_k.

Definition 2 (initial topic-based inspiration). *Let $\mathcal{G}(\mathcal{V},\mathcal{E})$ be a heterogenous information network, $\Delta\mathcal{T} = \{\Delta T_0, \ldots, \Delta T_N\}$ a set of time intervals and Θ a topic distribution. For a given topic z_k and a given threshold $\tau \in [0,1]$, 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. $ts_j \in \Delta T_0$, $ts_i \in \Delta T_1$ and $\theta_{jk} \geq \tau$.*

According to this definition, the initial inspiration takes place when an author a_l publishes a document d_j during ΔT_0 ($ts_j \in \Delta T_0$) covering topic z_k ($\theta_{jk} \geq \tau$), and another author a_h publishes a document d_i, during the following time interval

ΔT_1 ($ts_i \in \Delta T_1$). Notice that we do not impose any constraints on the topic covered by document d_i. Let us now introduce the definition of *subsequent topic-based inspiration.*

Definition 3 (subsequent topic-based 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. For a given topic z_k, 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 $v^a_h \xrightarrow{ad} v^d_i \xrightarrow{dd} v^d_j \xrightarrow{ad} v^a_l$ in $\mathcal{G}$ s.t. $ts_j \in \Delta T_{t-1}$, $ts_i \in \Delta T_t$ 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} for topic z_k.*

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 $\Delta T_t = [T_0 + t(\delta - \gamma), T_0 + t(\delta - \gamma) + \delta)$, for $t = 0, \ldots, N$ with $N = \left\lceil \frac{T_n - (T_0 + \delta - 1)}{\delta - \gamma} \right\rceil$.

3.2 Computation of the inspiration rank

We now describe how to assign a rank value to each topic depending on its inspiration speed. To this purpose, for a given topic z_k and a given set of time intervals $\Delta\mathcal{T} = \{\Delta T_0, \ldots, \Delta T_N\}$ we measure the cumulative number of new authors inspired at each time interval, according to the definitions of inspiration given in Section 3.1. In particular, given the heterogenous information network $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and a threshold τ, we call $A_0 = \{a_h | \exists (v^a_h, v^d_i) \in E^{ad} \wedge ts_i \in \Delta T_0 \wedge \theta_{ik} > \tau\}$ the set of authors that publish a paper on topic z_k during ΔT_0. Then, we define $A_1 = \{a_h \mid \exists a_l \in A_0 \;\; s.t. \;\; a_h \text{ is initially inspired by } a_l\}$ and, $\forall t = 2, \ldots, N$, $A_t = \{a_h \mid \exists a_l \in A_{t-1} \;\; s.t. \;\; a_h \text{ is subsequently inspired by } a_l \text{ during } \Delta T_t\}$. In a nutshell, A_1 is the set of initially inspired authors, $A_2, \ldots, A_N$ are the sets of subsequently inspired authors.

Finally, 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 $t = 2, \ldots, N$. We use this set to compute a linear function $y = \hat{\sigma}t + \hat{c}$ by solving the following simple linear regression problem

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

using the least squares method.

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 $\psi_k = \alpha + \sum_{w=1}^{N} \phi_{wk} n_w$
until $\frac{1}{K} \sum_k$ |change in $\psi_k| < \epsilon$;
return $\boldsymbol{\psi}$
```

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

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

where $\hat{\sigma}$ is the solution of the linear regression problem given in Equation 1.

Notice that, by varying parameters δ and γ, which define the width and overlap of time intervals in $\mathit{\Delta}\mathcal{T}$, different values of information rank can be obtained.

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

3.3 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 at describing 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 $\theta_d \sim \mathrm{Dirichlet}(\alpha)$, (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* [15]. In fact, traditional LDA implementations are based on either variational inference or collapsed Gibbs sampling; both methods require to process

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	6,959	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

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. The algorithm iteratively refines the variational parameters ϕ (line 4) that represents the word probability in each topic and ψ (line 5), which encodes the topics proportion within the document. When the procedure converges [15] ψ is returned as the topics assignment for the document represented by n (line 7).

4 Experiments

In this section, we present the results of our experiments conducted on a large corpus of scientific documents. In particular, we analyze the outcomes of our measure of inspiration in terms of effects on topic ranking. We compare our results with the standard diffusion approach, broadly adopted in most research works dealing with topic diffusion. In the following, we first describe the dataset used in our experiments and how we construct it; then, we provide some insights on the topic extraction and labeling tasks; finally, we give the details of the experimental protocol and report the results returned by our ranking-by-inspiration method in comparison with the standard ranking-by-diffusion approach.

4.1 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[23]: the *DBLP* and *ACM* citation networks[1]. The merge procedure is necessary because both datasets lack some in-

[1] `https://aminer.org/citation`

formation: 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[2].

4.2 Text processing and topic extraction

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 words list is also excluded from topic computation. First, documents are pre-processed with NLP techniques that perform tokenization, lemmatization, stop words 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 [19] that enables the extraction of an adaptive set of topics using an online learning version of Latent Dirichlet Allocation [15].

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.

Topics labeling. For improving the readability of our model, we introduce a simple topics 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

[2] 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 + ...	Artificial Neural Network, Artificial Neuron, -

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 Results

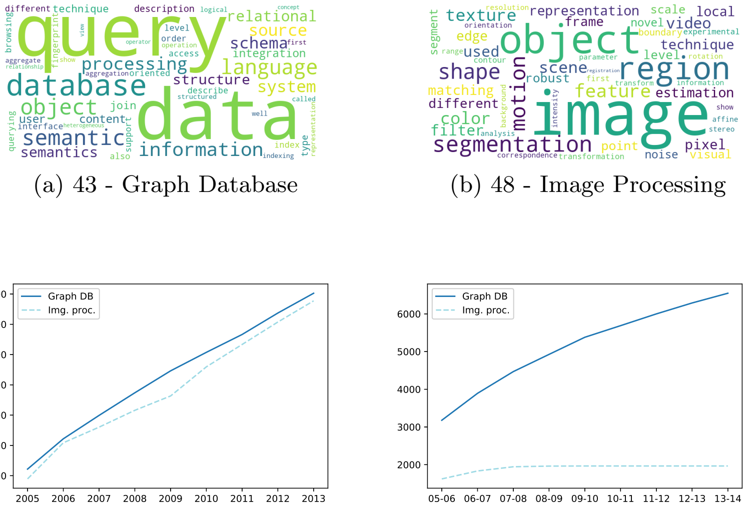
In our experiments, we calculate the ranking of topics according to their *inspiration rank* and diffusion rank in the time interval from 2005 to 2014 for $1 \leq \delta \leq 6$ and $0 \leq \gamma \leq \delta$. In all our experiments $\tau = 0.2$. Algorithms and scripts are implemented in Python, and data are stored in a MongoDB[3] database server. The source code and the dataset are available online[4]: the whole analysis process can be driven within an interactive Jupyter notebook[5]. The experiments are performed on a server with two 3.30GHz Intel Xeon E5-2643 CPUs, 128GB RAM, running Linux.

Examples of inspiration trend. As a first illustrative result, we show the inspiration trend of two topics, and compare it to their diffusion trend. The topics selected for this analysis are *Graph DB* (topic-43) and *Image Processing* (topic-48). They are intrinsically described in word clouds shown in Figure 2 by means of their more representative terms. These two topics have been selected due to their similarity in terms of number of assigned papers ($8,969$ for topic 43 and $8,646$ for topic 48), authors ($24,143$ for topic 43 and $23,056$ for topic 48) and distribution of papers in the considered time frame. According to Figure 2c, which shows the diffusion trend as computed by the method in [12], these topics have very close diffusion trends in the bibliographic network. However, there is a strong difference in the inspiration trend, as shown in Figure 2d: in fact, topic 43 (graph databases) evolves more rapidly than topic 48 (image processing). This behavior can be explained by the increasing and fast research results obtained by the database community, also boosted by the research on semantic queries and triplestores. Image processing, in contrast, appears as an evergreen albeit not particularly evolving research field in the time frame considered here. In this experiment, we employ $K = 50$, $\delta = 2$ and $\gamma = 1$.

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

[4] https://github.com/rupensa/tranet

[5] https://jupyter.org/



(a) 43 - Graph Database

(b) 48 - Image Processing

(c) Diffusion on network

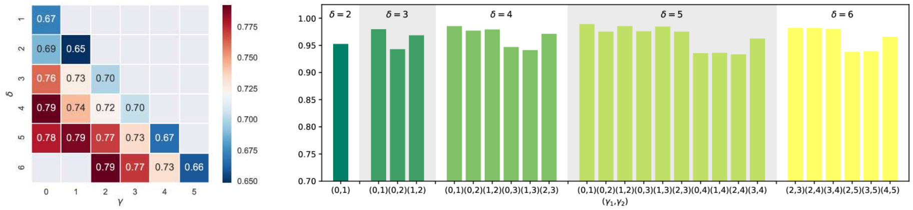
(d) Inspiration speed

Fig. 2: Diffusion and word clouds of the selected topics.

Inspiration rank vs. diffusion rank. In order to study the difference between the proposed ranking and the usual one, we measure the Spearman's rank correlation coefficient [22] between the the *inspiration rank* and *diffusion rank*. The Spearman's rank coefficient assesses monotonic relationships between two series of values. It basically captures the correlation between the two rankings and ranges between -1 (for inversely correlated sets of values) and $+1$ (for the maximum positive correlation).

Figure 3a shows the Spearman's rank correlation coefficient between inspiration rank IR and diffusion rank DR for several values of δ and γ. The empty tiles on the bottom left are due to lack of data: since our dataset covers only 10 years, when $\delta >> \gamma$ there is only one time interval valuable for calculating the rank, that is not sufficient for fitting a linear function.

In general, it can be noticed that the two ranks are always positively correlated. However, for lower values of δ (i.e., for small time windows), the correlation is sensibly slighter (only 0.67 for $\delta = 1$ and 0.65 for $\delta = 2$, with $\gamma = 1$). This can be explained by the fact that topics that diffuse faster are not necessarily the most inspiring ones, according to our definition. When the inspiration rank is high for small time windows, it means that citations occur very fast. The fact that the two rank values get more similar when δ increases, also confirms our intuition. In fact, it is more likely that papers are cited after four or five years, rather than the year following its publication. When this occurs, it means that this topic is evolving very fast, inspiring plenty of new research works. Another noticeable result is that correlation decreases when γ increases. This is due to the fact that larger overlap values allow to capture more citations to papers



(a) Spearman correlation values between inspiration and diffusion rank for several values of δ and γ.

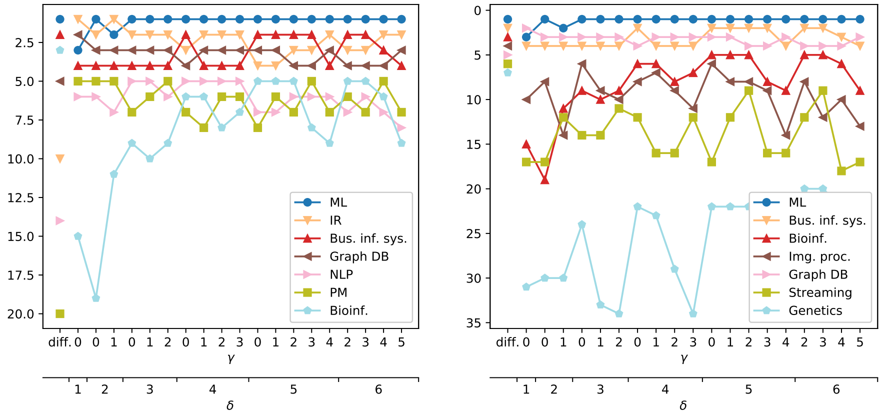
(b) Spearman correlation values of inspiration ranks computed with several γ values and same δ.

Fig. 3: Correlation computed between inspiration and diffusion ranks.

published in the previous interval. However, its effect is weaker than the one of parameter δ, as shown in Figure 3b, where the correlation between any pair of γ values for the same value of δ are illustrated. This particular results also shows that our method is rather stable toward variations of parameter γ.

Ranking comparison. Here, we analyze the ranking of the top 7 topics based on the average inspiration rank, compared to the ranking of the same topics based on diffusion rank. The results are depicted in Figure 4a (notice that diffusion ranks are not affected by parameter variation). We notice that the best 4 topics are almost the same ones for all values of δ and γ, then the ranking becomes more chaotic. More interestingly, topic IR (Information retrieval) is always ranked in the top 3 positions for inspiration, while it is ranked $10th$ according to diffusion. Our measure capture 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 DB (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 PM (pattern mining) among the top 7 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. These topics are only ranked 13th and 20th according to standard diffusion metrics.

It is worth noting that, by analyzing the ranking of the top 7 topics based on the average diffusion rank, and their respective ranking based on inspiration (Figure 4b), 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 third) has spread rapidly in the last 10 years. However, in our approach this topics gets a lower rank: this can be explained by the fact that, in the



(a) Top 7 topics by inspiration

(b) Top 7 topics by diffusion

Fig. 4: Ranking by inspiration and diffusion for several values of δ and γ. Diffusion rank are not affected by parameter variation.

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 genetics and image processing.

Finally, we explore the ranking provided by the two methods for a set of topics of interest, namely: Deep learning (topic-3), Clustering (topic-22), Information retrieval (topic-26), Neural networks (topic-33) and Pattern mining (topic-41). The ranking positions are shown for several values of δ and γ in Figure 5. The same conclusions drawn in the previous experiments hold here, in particular for pattern mining (PM). Interestingly, deep learning is ranked low, despite its objective success in the last five years. This may be explained by the fact that the topic model has been trained on papers published from 2000 to 2004 when deep learning was beginning to be recognized as a research field itself. Furthermore, since we only consider Computer Science venues, the broad influence of deep learning on other research areas can not be captured. Notice that, however, it is always ranked far higher by our method. This is another result indicating that inspiration capture a more realistic influence semantics than simple topic diffusion.

5 Conclusions

We have proposed a new definition of influence that takes into account the inspiration of a given topic within a citation network. We have defined a new influence measure, called *inspiration rank*, that captures the inspiration rate of topics extracted by an adaptive LDA technique, within a given time interval.

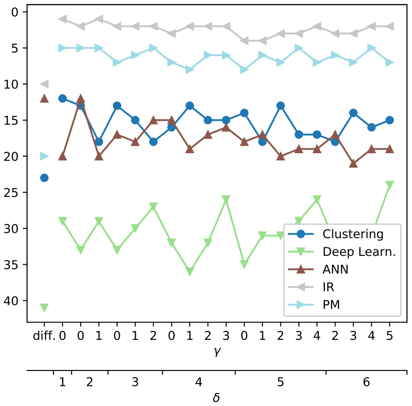


Fig. 5: Rankings on inspiration and diffusion speeds for several values of δ and γ on a set of selected topics.

The *inspiration rank* allows the discovery of the most inspiring topics according to different levels of speed. We have shown experimentally the effectiveness of our measure in detecting the most inspiring topics in a citation network built upon a large bibliographic dataset. Although the core application is the analysis of topic 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.

As future work, we will define new author and paper ranking methods based on our inspiration measure. Furthermore, we will investigate new algorithms to learn the topic diffusion parameters under different diffusion models adopting our definition of inspiration.

Acknowledgments. This work is partially funded by project MIMOSA (MultI-Modal Ontology-driven query system for the heterogeneous data of a SmArtcity, "Progetto di Ateneo Torino_call2014_L2_157", 2015-17).

References

1. Bakshy, E., Rosenn, I., Marlow, C., Adamic, L.A.: The role of social networks in information diffusion. In: Proceedings of WWW 2012. pp. 519–528. ACM (2012)
2. Barbieri, N., Bonchi, F., Manco, G.: Topic-aware social influence propagation models. Knowl. Inf. Syst. 37(3), 555–584 (2013)
3. Boguslawski, B., Sarhan, H., Heitzmann, F., Seguin, F., Thuries, S., Billoint, O., Clermidy, F.: Compact interconnect approach for networks of neural cliques using 3d technology. In: Proceedings of IFIP/IEEE VLSI-SoC 2015. pp. 116–121 (2015)
4. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of ACM SIGKDD 2009. pp. 199–208. ACM (2009)

5. Coates, A., Huval, B., Wang, T., Wu, D.J., Catanzaro, B., Ng, A.Y.: Deep learning with COTS HPC systems. In: Proceedings of ICML 2013. pp. 1337–1345. JMLR.org (2013)
6. Cui, P., Wang, F., Liu, S., Ou, M., Yang, S., Sun, L.: Who should share what?: item-level social influence prediction for users and posts ranking. In: Proceeding of ACM SIGIR 2011. pp. 185–194. ACM (2011)
7. Daley, D.J., Kendall, D.G.: Epidemics and rumours. Nature 208, 1118 (1964)
8. Gohr, A., Hinneburg, A., Schult, R., Spiliopoulou, M.: Topic evolution in a stream of documents. In: Proceedings of SIAM SDM 2009. pp. 859–870. SIAM (2009)
9. Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters 12(3), 211–223 (2001)
10. Gruhl, D., Guha, R.V., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. In: Proceedings of WWW 2004. pp. 491–501. ACM (2004)
11. Gruhl, D., Liben-Nowell, D., Guha, R.V., Tomkins, A.: Information diffusion through blogspace. SIGKDD Explorations 6(2), 43–52 (2004)
12. Gui, H., Sun, Y., Han, J., Brova, G.: Modeling topic diffusion in multi-relational bibliographic information networks. In: Proceedings of CIKM 2014. pp. 649–658. ACM (2014)
13. He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., Giles, C.L.: Detecting topic evolution in scientific literature: how can citations help? In: Proceedings of ACM CIKM 2009. pp. 957–966. ACM (2009)
14. Hethcote, H.W.: The mathematics of infectious diseases. SIAM Review 42(4), 599–653 (2000)
15. Hoffman, M.D., Blei, D.M., Bach, F.R.: Online learning for latent dirichlet allocation. In: Proceedings of NIPS 2010. pp. 856–864 (2010)
16. Kempe, D., Kleinberg, J.M., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of ACM SIGKDD 2003. pp. 137–146. ACM (2003)
17. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. TWEB 1(1), 5 (2007)
18. Radicchi, F., Fortunato, S., Markines, B., Vespignani, A.: Diffusion of scientific credits and the ranking of scientists. Phys. Rev. E 80, 056103 (2009)
19. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50 (2010)
20. Seo, J., Seok, M.: Digital CMOS neuromorphic processor design featuring unsupervised online learning. In: Proceedings of IFIP/IEEE VLSI-SoC 2015. pp. 49–51. IEEE (2015)
21. Shi, X., Tseng, B.L., Adamic, L.A.: Information diffusion in computer science citation networks. In: Proceedings of ICWSM 2009. The AAAI Press (2009)
22. Spearman, C.: The proof and measurement of association between two things. The American Journal of Psychology 15(1), 72–101 (1904)
23. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: Extraction and mining of academic social networks. In: Proceedings of KDD 2008. pp. 990–998 (2008)
24. Yang, J., Counts, S.: Comparing information diffusion structure in weblogs and microblogs. In: Proceedings of ICWSM 2010. The AAAI Press (2010)
25. Yang, J., Counts, S.: Predicting the speed, scale, and range of information diffusion in twitter. In: Proceedings of ICWSM 2010. The AAAI Press (2010)