Complex Systems Informatics and Modeling Quarterly (CSIMQ)
eISSN: 2255-9922
Published online by RTU Press, https://csimq-journals.rtu.lv
Article 98, Issue 17, December 2018/January 2019, Pages 24–43
https://doi.org/10.7250/csimq.2018-17.02

Time-Dependent Influence Measurement in Citation Networks

Monika Ewa Rakoczy[1]*, Amel Bouzeghoub[1], Alda Lopes Gancarski[1], and Katarzyna Wegrzyn-Wolska[2]

[1]SAMOVAR, CNRS, Telecom SudParis, 9 Rue Charles Fourier, Evry, France
[2]Efrei Paris, 30 Avenue de la Republique, 94800 Villejuif, France

monika.rakoczy@telecom-sudparis.eu, amel.bouzeghoub@telecom-sudparis.eu,
alda.gancarski@telecom-sudparis.eu, katarzyna.wegrzyn@groupe-efrei.fr

Abstract. In every scientific discipline, researchers face two common dilemmas: where to find bleeding-edge papers and where to publish their own articles. We propose to answer these questions by looking at the influence between communities, e.g. conferences or journals. The influential conferences are those which papers are heavily cited by other conferences, i.e. they are visible, significant and inspiring. For the task of finding such influential places-to-publish, we introduce a Running Influence model that aims to discover pairwise influence between communities and evaluate the overall influence of each considered community. We have taken into consideration time aspects such as intensity of papers citations over time and difference of conferences starting years. The community influence analysis is tested on real-world data of Computer Science conferences.

Keywords: Influence, Influence Estimation, Citation Networks, Social Networks, Granger causality.

1 Introduction

Discovery, detection, and evaluation of different social behaviors, such as trust, influence, or reputation in on-line social networks have drawn much focus in current research [1]. In particular, the investigation of influence has been a recent interest in research, as it is especially useful in real-life applications such as political campaigns, marketing strategies or recommending products.

Due to the fact that the notion influence is complicated, the attempts at studying influence aim at different aspects and properties of the term. There are few well-known algorithms, such as degree centrality, or PageRank, that were focused on exploiting only the topological structure of social networks, and thus identifying the central role and high connectivity in the network structure with influence. Other studies have brought a wider range of influence aspects [2], such as a focus on the nature of the interaction between users, strength, frequency of these interactions, etc. The research on influence has led to various studies examining various influence aspects, e.g. analysis of global level influence (i.e. the impact of a user on the whole network) or reciprocal

* Corresponding author

Reference: M. E. Rakoczy, A. Bouzeghoub, A. L. Gancarski, and K. Wegrzyn-Wolska, “Time-Dependent Influence Measurement in Citation Networks,” Complex Systems Informatics and Modeling Quarterly, CSIMQ, no. 17, pp. 24–43, 2018. Available: https://doi.org/10.7250/csimq.2018-17.02

Additional information. Author’s ORCID iD: M. E. Rakoczy – https://orcid.org/0000-0001-5249-5509. PII S225599221800098X. Received: 14 October 2018. Accepted: 27 November 2018. Available online: 31 December 2018.

influence between users (i.e. individual impact of one user towards another) [3]; influence discovery depending on topic [4]; examination of influence propagation and aggregation [5]; and investigation of consequent influence path and its direct or indirect effect [4]. However, due to the complexity of influence, the notion is still to be explored. Many of existing works focus the scope of the studies on influence identification solely for individuals – influencers, and studying their impact on network structures. Moreover, the influential node discovery algorithms often focus on producing only overall ranking, by which it is possible to order all the nodes from having the most to the least influence. Nevertheless, the information about particular influence relations, i.e. whether a particular user is influencing another one is lost. Our focus is to study those pairwise relations, in order to gain better knowledge about network dependencies and possibly have more insight into influence propagation between particular entities in the network. Furthermore, ranking methods, such as the popular PageRank [6], have underlined assumption that the relations between nodes from which the social graph emerges are already given and defined as the influence. This leads to simplifying the notion of influence, by defining it as accessible relations between nodes such as friendships, tweet ratio, etc. Following [1], we argue that the influence discovery must rely not only on relations between nodes but also on the activity dependence or causality between nodes.

The second major focus of this article is targeted on the time-dependency of the networks. While the need of including time constraint in the analysis of social networks is undeniable, having applications including discovering voting and trading patterns, voice calls tracing [7], etc., it is a difficult and complex problem due to changing network patterns. Importantly, the type of the network when considering time-dependency is not to be omitted. Depending on the particular network, the adopted assumptions about the time can be vastly different. For instance, Twitter can be characterized by short-living messages and quickly expiring influence [7]. On the other hand, blogs have less dynamic characteristics, where a particular post can gain recognition much slower, in comparison to tweets. In this article, we precisely focus on citation networks scope, in which the impact of particular paper is time-sensitive and insight may change during the lifetime of the network, even if it tends to slowly fade rather than rapidly disappear.

Finally, it is important to mention also the role of communities in social networks, which are known to have different structural patterns and tend to propagate information differently than individuals. While community detection is out of the scope of this article, we do use predefined communities, in order to observe and study influence dependencies in the network.

This article is an extension of our previous work from [8]. Here, we further extend the model of pairwise influence between predefined communities, called Running Influence (RI) which allows obtaining information about influenced entities, and select the influenced community subset and finally evaluates the impact that communities have on each other. We propose an additional metric for evaluating impact communities have on each other within citation networks, Reference Ratio, as an addition to the Time-Dependent Citation Ratio. We also show extensive experiments using real-world dataset, which enable us to compare proposed metrics and study the impact of communities.

The remainder of the article is structured as follows. Section 2 describes state-of-the-art dealing with influence, communities and citation networks. Section 3 introduces preliminary notions needed and specifies our research problem in detail. The metrics for communities impact, Citation Ratio and Reference Ratio, along with the model of Running Influence are described in Section 4. Section 5 includes the description of real-world Microsoft Academic database used for experiments, as well as discusses the results of the experiments. Finally, we summarize the work and introduce some directions for future research in Section 6.

2 Related Works

In this section, we discuss some of the state-of-the-art works connected to the subject of influence. We first briefly summarize influence metrics. Then, we introduce a few works focusing on identifying influence between communities. Finally, we show some works dealing with citation networks, with the special regard towards the scope of this article. In the light of these considerations, we emphasize the research questions that we aim to address.

2.1 Influence Metrics

The most popular and known methods for influence detection and estimation are the ones focused on the typology of the network. Methods such as PageRank, betweenness or degree centrality or HITS [9] quantify the influence of every node and, as a result, return a rank of highest to lowest influencers, that is based on a particular metric evaluating structural properties of the graph. Other approaches, based on social media sites were also proposed, which for influence model considered also factors such as tie strength between users [10], or influencer's local or global environment [11]. There were also studies dealing with influence in broader perspective, e.g. dealing with influence propagation in social networks and the problem of maximization of this propagation [5], or studying of the implications of social influence on the network evolution [12]. However, contrary to our work, while these works focus on creating a hierarchy containing each individual, they do not aim to discover and evaluate the point-to-point influence that is targeted at communities. Ranking methods also base the calculation on the notion that connections between nodes are already an influence (i.e. graph is representing influence dependence) and thus reducing the notion only to topological features of the term. Therefore, they omit the process of the influence discovery, which is crucial for extracting complexity of influence.

2.2 Communities

Some research has been done for detection and evaluation of influence within communities. Authors in [13] proposed detecting influence of communities, however, they focused on information propagation flow and omitted the problem of influence discovery, basing their work on assumption of a similar influence tendency. The work [14] analyzed influence in social groups within content-sharing based social networks, with an examination of the process of joining of users. Similarly, work in [15] also focused on detecting groups of users with common interests rather than studying their influence. [16] proposed a model of influence based on topic, investigating both direct and indirect influence, but not dealing with communities. On the other hand, the work in [17] studied citation networks, nevertheless, it targeted the tracking of the influencer-influencee relationships.

Works such as [18], [19], and [20] study the dynamicity of networks connected to the evolution of the communities, i.e. examination of the membership of each node to particular clusters and how these memberships change in time. Some approaches deal with community detection at each snapshot, some include composition of historic and current information about the graph in order to determine and track the communities. Nevertheless, these methods, while focusing on detection of the communities and their evolution, are not suitable for tracking precisely influence dependencies for explicitly defined communities. Furthermore, as both influence is highly context-specific [9] and different network types present various dynamicity characteristics, time-dependency models are not universal. Work [7] creates time-respecting dynamic approach for capturing influence of Twitter, however highly time-sensitive, quickly decaying influence of a tweet differs greatly from influence concerning citation between scientific papers. There have been several research works dealing with issues within the citation networks scope, such as studying papers topic evolution [21], predicting paper influence [22] or ranking experts [23], however, not focusing neither on pairwise influence between communities nor on problem of time-sensitiveness of influence.

2.3 Citation Networks

An important measure in the context of citation network is the famous work of Hirsch [24]. The author presents a metric that aims to calculate the influence of a researcher, based on his/her publications. While it is widely popular, it can be manipulated via self-citations. Moreover, it does not consider the causality between the citations. The problem we are studying is close to the work shown in [25]. This work is introducing the model of group influence which uses Granger causality concept [26] to determine the direction of pairwise influence between two communities, understood as two conferences, for a particular time period. However, the model is capable only of returning pairs of dependent conferences. Therefore, we cannot draw any further conclusions about influence dependencies between the conferences, as there is no measure for quantifying the community influence.

In this work, we particularly focus on the issue of: (1) how to measure the impact between communities considering its time-dependent characteristics, (2) how to quantify and evaluate pairwise influence in order to compare communities influence, (3) how in citation networks we can model time decay.

3 Preliminaries

In this section, we introduce the notion of citation graph and a closely related community-based citation graph; describe formally the problem definition, and, lastly, discuss the influence properties that we then target in our model.

A citation network can be depicted as a graph, where nodes symbolize papers and edges are the citations between them. In order to consider time for such *citation graph*, we can model the dynamic citation network as a sequence of directed graphs $\{G_1, ..., G_m\}$ for time $\{t_1, ..., t_m\}$, where G_i represents a directed citation graph at particular snapshot t_i. G_i is thus defined as a directed graph $G_i = \{V_i, E_i\}$, where V_i signifies set of papers (nodes), and E_i describes the set of citations (edges) between papers, from citing to referenced paper.

Basing on this classic citation graph, we will define a *community-based citation network*, in order to better illustrate the focus of this work. As mentioned before, in this study we target influence in the dynamic citation networks between predefined communities. Therefore, first, we define the constant set of of-interest communities $S = \{C_1, ..., C_n\}$, with assigned set of papers and citations. Similarly to above-mentioned citation graph, we model the dynamic community-based citation network as a sequence of directed graphs $\{CG_1, ..., CG_m\}$ for time $\{t_1, ..., t_m\}$, where CG_i represents a directed community-based citation graph at particular snapshot t_i. Importantly, due to dynamic nature of the graph, while the set of of-interest communities is constant (these are the communities whose influence we study at each snapshot t_i), in each graph CG_i there can be other appearing or disappearing communities and citations between them at each snapshot t_i. CG_i is thus defined as a directed graph $CG_i = \{CV_i, CE_i\}$, where CV_i represents all communities at time t_i, and CE_i signifies citations between communities at time t_i, leading from citing community to cited community with associated weight above each edge symbolizing the number of citations at time t_i. In other words, the community-based citation graph is an aggregated version of citation graph, where papers are combined together to form nodes-conferences, and the aggregated citations between conferences, i.e. group of papers are the vertices. An example of such community-based citation graph is shown in Figure 1. While the set of predefined communities $S = \{C_1, ..., C_n\}, S \subseteq CV_i$ can be the result of any static community mining algorithm, in our work we assume that this set is created using the venue of each paper, therefore each community C_i corresponds to a conference (hence, set of papers published at particular conference C_i). Due to the fact that in this work we use only the community-based citation graphs, we will reference it simply as citation graph or citation networks.

Problem definition For time snapshots $\{t_1, ..., t_m\}$, given the universe of conferences U, the set of of-interest, predefined communities $S = \{C_1, ..., C_n\}(S \subset U)$, and the number of citations between each conference pair at particular snapshot of time $t_i \in \{t_1, ..., t_m\}$, our goal is threefold: (1) for each pair of conferences to determine pairwise Running Influence (RI) for time period

$[t_s, t_e] \subset \{t_1, ..., t_m\}$ using citation information from each time $t_i \in [t_s, t_e]$; (2) to create RI graph (influence dependency graph) for selected set of communities; and (3) to estimate overall value of the RI for time period $[t_s, t_e] \subset \{t_1, ..., t_m\}$ for all considered communities from S.

While the issue of influence measuring is a complex problem, we also want to focus on capturing in our model important properties of influence. In particular, the influence has to have the following features:

- asymmetry,
- direction,
- time-dependency.

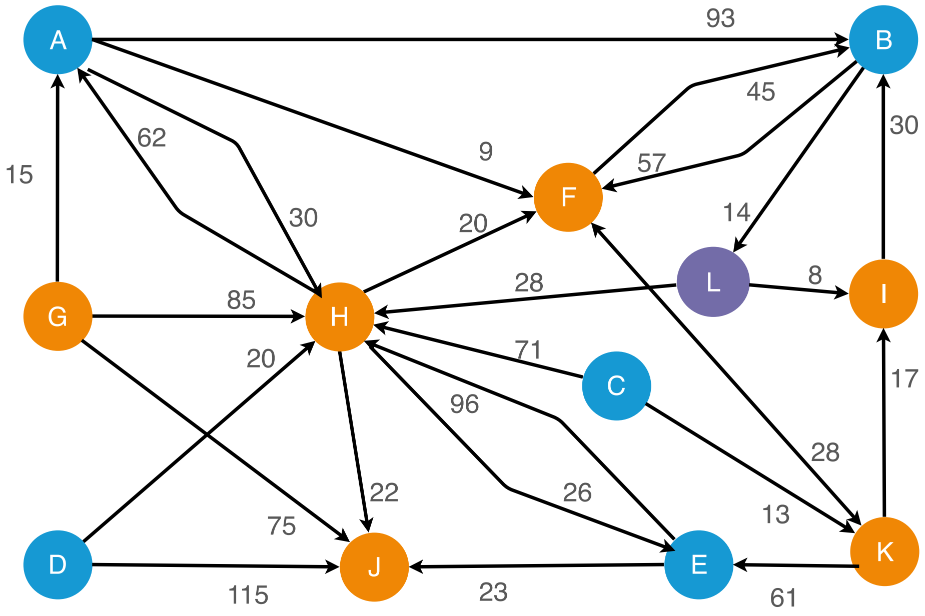


Figure 1. Example of a community-based citation graph containing references between conferences at time t from conference universe $U = \{A, B, C, D, E, F, G, H, I, J, K, L\}$, where each vertex symbolizes conference, each edge symbolizes reference, with the number of citations indicated above the edge, and colors indicate different conference topics.

4 Proposed Model

In order to detect and evaluate influence, according to properties presented in the previous section, we divided the process of influence discovery and assessment into three steps:

1. Evaluating the effect of one conference onto another, for each pair of conferences A, B, which belong to a predefined set of of-interest communities S for each time snapshot $t_i \in [t_s, t_e]$, using one of two different measures (Section 4.1):
   (a) utilizing Citation Ratio
   (b) using Reference Ratio
2. Determination of Running Influence RI for each pair of conferences from conference set S and direction of this influence within the time period $[t_s, t_e]$ using Citation Ratio (Section 4.2),
3. Evaluating overall Running Influence value between communities from set S, using the citation ratio series at each time snapshot $t_i \in [t_s, t_e]$, with consideration of influence direction (Section 4.3).

We discuss each step in the following subsections.

4.1 Measurements for Effects Between Communities

Before assessing the Running Influence between two conferences in citation networks at a given period of time, we need to calculate the effect one community has on the other in particular snapshot t_i. Here, we show two possible measures to evaluate it. We first discuss the Citation Ratio extension, a measure based on the work [27], [28] enhanced in order to capture time-dependence of the network. Then, we introduce a second measure, Reference Ratio.

Citation Ratio To capture the effect one community has on another in particular snapshot t_i we have used the metric based on normalized citation weight [27], [28]. This metric measures a number of citations made by one community to the other. Citation Ratio for community A towards community B at time t is defined as:

$$CR_{A\to B}(t) = \frac{|cit_{B\Rightarrow A}(t)|}{\sum\limits_{i=1}^{|S|} |cit_{B\Rightarrow i}(t)|} \tag{1}$$

where $|cit_{B\Rightarrow A}(t)|$ is the number of citations at time t from conference B that reference papers of conference A (created until time t) and the denominator of the formula is the aggregate number of all the citations at time t made by conference B to any conference paper (created until time t) in the set S (including conference A). S symbolizes a group of conferences which we consider as a "general universe of conferences" to which conference B can have references while calculating Citation Ratio, which significance we will now detail. The arrow in $CR_{A\to B}$ indicates the direction of the citation influence. Moreover, contrary to previous works [25], which ignore the size of the considered possible reference group, we propose to distinguish two interpretations of Citation Ratio metric, namely local and global one. While a detailed discussion on the two types can be found in our previous work [8], we will only limit ourselves to recall the two definitions.

Definition 1 ***Global Citation Ratio*** *is the ratio given by Formula (1), where in the denominator we consider all possible communities from available data (set S is equal to universe U). The result of this metric, assuming that we have a wide range of conferences in the dataset, captures the ratio between citations from a particular conference to all conferences in a global and unbiased way.*

Moreover, the measure of Global Citation Ratio also captures one of the dynamic features of the network: over different time snapshots conferences can appear and disappear (independently from the relation between two conferences that the CR is calculated). Thus, using the maximal possible set of conferences at every snapshot leads to include this aspect of dynamicity of the network.

Definition 2 ***Local Citation Ratio*** *is the ratio given by Formula (1), where in the denominator we consider a subset of all available data. The result of this metric captures the impact of one conference on another in a deliberately created subset, where the number of conferences that can be referenced is limited. It can be used in order to obtain the relative, scope measure of the impact between conferences.*

Selection of the subset S of universe U of all conferences can be done in any way which results in a subgroup of conferences. It is worth noticing that the selection of this subset can lead to the bias in the Local Citation Ratio in case where the set of selected conferences will contain two smaller subsets between which there will be no citations. In this case, the Local Citation Ratio can be artificially high for a particular pair of conferences, since none of other in the set is connected to them. For instance, in order to select subset S for the use of Local Citation Ratio a subgroup of same topic conferences can be used. Resulting Local Citation Ratio will describe the impact between two conferences within the group of same scope conferences (which creates subgraph of closely connected communities).

Time-dependent Citation Ratio While Citation Ratio is calculated at each time snapshot, it is important to notice that in the basic version, the metric (Formula (1)) treats equally all the citations made in time t to publications published at any given time (obviously, before time t). However, citations are time-sensitive – indeed, it would be expected for more recent papers to be cited more often. On the other hand, they are not very rapidly disappearing, in comparison to, for instance, tweets [7]. It is also important to notice that some of the considered pairs of conferences may not start operating at the same time (i.e. year). Let us assume two conferences A and B for which we want to calculate $CR_{A\to B}(t_i)$. At time t_i, conference B can cite papers from conference A published in the past until time t_i. In order to study the impact of time on the citations, we aim to differentiate the impact of citations of articles published in each year $\{t_j\}$ before time t_i ($t_j < t_i$) on the $CR_{A\to B}(t_i)$ value. To do so, we propose (1) a weight function pf, that prioritizes citation information for particular points in time, (2) to calculate the $CR_{A\to B}(t_i)$ value as series of the ratios, where we divide the citations made by conference B by the time of publication of the article to which the citation is made. Each such partial series is then multiplied by the weight from above-mentioned weight function. The proposed method is described in Algorithm 1.

Algorithm 1 Method for calculating Citation Ratio value $CR_{A\to B}(T)$ for a particular time ($computeCR$) with consideration of time-dependency

```
Input: Conference set $S \subset U$ for which we consider influence, conferences $A$ and $B$ for which
    we calculate $CR_{A\to B}(T)$, max time $T$, $t_A$ start of conference $A$, priority function $pf$, time step
    $y$
Output: $CR_{A\to B}(T)$ value for $T$ prioritized according to priority function $pf$
1: $CRT\_series_T = 0$
2: for $t_i \Leftarrow t_A$ to $T$ with step $y$ do
3:     $CRT\_series_T + = \frac{cit_{B\to A}(t_A,t_i)}{\sum_{i=1}^{|S|} |cit_{B\to i}(t_A,t_i)|} \times pf(t_i, t_A, T)$
4: end for
5: $CR_{A\to B}(T) = sum(CRT\_series_T)$
```

For each time t_i, Algorithm 1 calculates the number of citations of conference B made to papers of A published in time t_i, divides it by the total number of citations of conference B to any conference from subset S, and multiplies this ratio by the weight assigned according to the particular time t_i (lines $1-3$). The priority function pf is any real-valued, well-defined function over domain $[0, 1]$. It changes the importance of historical citations over time. Since we operate on time snapshots, we convert a point in time to [0, 1] domain. For instance, a constant priority function emphasizes equally information in each time snapshot. Finally, the result $CR_{A\to B}(T)$ is obtained by aggregating partial results (function sum, line 4). In order to obtain Citation Ratio series for a particular time period, the algorithm has to be repeated for each time point within the considered time period.

Reference Ratio Apart from Citation Ratio, we also introduce a second measure for capturing the effect one community has on another within particular time snapshot t_i, called Reference Ratio. Reference Ratio for community A towards community B at time t is defined as:

$$RR_{A\to B}(t) = \frac{|cit_{B\Rightarrow A}(t)|}{\sum_{i=1}^{|S|} |cit_{i\Rightarrow A}(t)|} \tag{2}$$

where $|cit_{B\Rightarrow A}(t)|$ is the number of citations at time t from conference B that reference papers of conference A (created until time t), the denominator of the formula signifies the total number of all references at time t made to papers of conference A (created until time t) by any conference in the set S. Similarly to Citation Ratio, S symbolizes a group of conferences which we consider as a ”general universe of conferences” to which conference B can have references while calculating Reference Ratio. The arrow in $RR_{A \to B}$ indicates the direction of the citation influence. Comparably to Citation Ratio, depending on the considered set of communities S, we can distinguish ***Local Reference Ratio*** and ***Global Reference Ratio***.

Similarly to Citation Ratio, we consider a time-dependent version of Reference Ratio, by utilizing the priority function. Since the method is basically the same, we do not include the pseudo-code of the algorithm. Notably, the only change needed is replacing formula for Citation Ratio with the formula from Equation 2 of Reference Ratio. The difference between Citation Ratio and Reference Ratio can be seen in Figure 2. While Citation Ratio measures how much citations of conference B (in time t) were dominated by citation to the selected conference A, Reference Ratio is actually checking how much the group of citations from B to A is dominating all the references to conference A in time t. It can be seen that while both ratios are measuring impact between two conferences, the implications of those two metrics are very different.

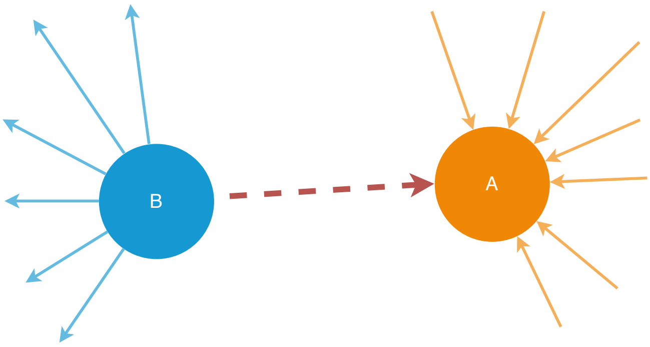


Figure 2. Differentiation of CR and RR metrics in time snapshot: CR considers in denominator all B citations including citation to A (blue arrows and red arrow), whereas RR denominator considers all citations to A (orange arrows and red arrow)

4.2 Running Influence Determination

In order to determine the Running Influence between communities and its direction, we incorporated the Granger causality into our model, similarly to the model in [25]. Thanks to using Granger causality on the time series of Citation Ratios, we capture the causality notion between those series. This broadens the sense of influence between two conferences, which is based not only in typology (from Citation Ratio component) but also includes the notion of dependence of action (i.e. citing) between conferences. This leads to improved Running Influence model in terms of hidden knowledge discovery. Before the presentation of the RI model, we briefly introduce the Granger causality method in what follows.

Granger Causality Intuitively, Granger causality can be explained [26] as: *Y(t) is causing X(t) if we are better able to predict X(t) using the history information of both X(t) and Y(t) than solely using the history information of only X(t)*, where X and Y are two stationary series X and Y, defined as:

$$\begin{aligned} X(t) &= \sum_{j=1}^{m} a_j X(t-j) + \sum_{j=1}^{m} b_j Y(t-j) + \zeta(t) \\ Y(t) &= \sum_{j=1}^{m} c_j X(t-j) + \sum_{j=1}^{m} d_j Y(t-j) + \eta(t) \end{aligned} \tag{3}$$

where m is the maximal time lag, and it is assumed to be finite and shorter than given time series; a_j and c_j are the matrices containing the coefficients of the model; ζ and η are prediction errors (uncorrelated white noise series).

In order to determine Granger causality between two series of data, for a particular time period, Granger causality test has to be performed. There are different Granger causality tests proposed in the literature, such as Sims test, Wald test, Lagrange-multiplier test [29]. The test checks if the hypothesis of no Granger causality between two series (so-called null hypothesis) is true. We say that X granger causes Y if the null hypothesis is rejected, meaning $b_j \neq 0$; otherwise, the null hypothesis holds, hence X do not granger causes Y. Similarly, we say that Y granger causes X if $c_j \neq 0$ (null hypothesis is rejected); otherwise Y do not granger causes X. It is important to notice that we can observe bidirectional granger causality if both X granger causes Y and Y granger causes X. The result of the Granger causality test is the P-value which, depending on assumed threshold ϵ, determines if null hypothesis holds (if P-value is smaller than ϵ) or is rejected (if P-value is greater than ϵ). The ϵ value is arbitrary given, depending on how significant the results should be.

Pairwise Running influence Presented above Granger causality is used in order to determine the causal relation between two time series obtained via calculating Citation or Reference Ratio for a particular time period. In particular, we determine Granger causality relation between a pair of conferences, e.g. A and B, using one of the two ratio metrics for obtaining time series. We calculate CR (or RR) for each time snapshot t_i from a particular time period T for which we determine RI ($t_i \in T$), from A to B and from B to A, so that we have two separate CR time series: (1) CR series (RR series) of conference A impacting B (i.e. $Y(t)$ in formula 3) and (2) CR series (RR series) of conference B impacting A (i.e. $X(t)$ in formula 3). Next, we determine Granger causality of A towards B accordingly to the method described above. The result of the Granger causality test answers the question of whether conference A influences conference B, hence determines the possible direction of running influence for a particular time period. We can formally define Running Influence as follows:

Definition 3 ***Running Influence*** *of community A towards community B, indicated $RI_{A \to B}$, for time period $[t_s, t_e]$, can be observed when Citation Ratio series between time t_s and t_e of conference A towards B Granger causes Citation Ratio series between time t_s and t_e of conference B towards A, with the level of significance ϵ in Granger causality test. Moreover, the time period $[t_s, t_e]$ has to respect the following conditions:*

- *Start of the time period t_s for which influence is measured has to be equal to or after the start moment of both conferences A, B, i.e. $t_s \geq t\_beg_A \wedge t_s \geq t\_beg_B$,*
- *End of the time period t_e has to be equal to or before the moment that both conferences A, B seize to exist, i.e. $t_e \leq t\_end_A \wedge t_e \leq t\_end_B$.*

The process of determination if one conference influences another can be seen in Algorithm 2. The first for loop (lines $1-4$) creates CR series (or RR series depending on the formula used) for each time snapshots for the two considered conferences. Then, in lines $5-13$ Granger causality test is performed, and the final existence of RI (or its lack) is determined (lines $9-13$).

In our experiments, we assume the value of ϵ equal to 0.01. Additionally, as it can be observed in Granger causality definition, the Running Influence can be bi-directional, in a case when both considered conferences have an impact on one another in the same time period.

4.3 Overall Running Influence Estimation

Running Influence described in Subsection 4.2 is used to determine the existence and direction of the running influence between two communities within the same time period $[t_s, t_e]$. However, it is not sufficient; indeed, while just detecting pairwise influence between communities can be useful, the metric evaluating the value of the Running Influence of a particular conference is still needed.

Algorithm 2 Method for determining Running Influence from conference A to B – whether conference A influences conference B

Input: Two conferences A and B for which Running Influence is determined, period $[t_s, t_e]$, maximal lag number m, significance level ϵ, time window step y, pf priority function

Output: Running Influence direction (or lack of RI) from conference A to B during time period $[t_s, t_e]$

```
 1: for $t_i \Leftarrow t_s$ to $t_e$ with step $y$ do
 2:     $CR\_series_{A \to B} + = computeCR(A, B, t_i, pf)$
 3:     $CR\_series_{B \to A} + = computeCR(B, A, t_i, pf)$
 4: end for
 5: for $lag \Leftarrow 1$ to $m$ do
 6:     $GCResult_{B \to A} + = GCTest(CR\_series_{B \to A}, CR\_series_{A \to B}, lag)$
 7: end for
 8: $minGCResult_{B \to A} \Leftarrow min(GCResult_{B \to A})$
 9: if $minGCResult_{B \to A} < \epsilon$ then
10:     $RI_{A \to B} = true$
11: else
12:     $RI_{A \to B} = false$
13: end if
```

In order to calculate the value of the Running Influence during time period $[t_s, t_e]$, we propose Algorithm 3.

Algorithm 3 first calculates Citation Ratio (or, alternatively Reference Ratio) for every snapshot of time, creating Citation Ratio series using function $computeCR$ (Reference Ratio series using function $computeRR$), for each pair consisting of conference A for which the influence value is calculated and conference from subset S of which we consider influence (lines $2 - 5$). Then, a set of Granger causality tests are performed (lines $6 - 9$), each with different lag value, which defines how many past values should be taken into consideration. Set of $GCTest$ as a result returns set of P-values, by which we can determine the existence of influence by selecting the minimal P-value from the $GCResult$ variable (function min in 9th line) and checking if it is smaller than assumed level of significance ϵ (lines $10 - 12$). Since we iterate (line $1 - 13$) over the set of conferences S (excluding conference A) performing the sequence of actions described above, we afterwards obtain subset $InfSet \subset S$, which contains all conferences influenced by conference A. This process concerns determining Running Influence from A to any other considered conference from S (pairwise RI), and was explained in detail in Section 4.2.

Then, we estimate the Running Influence value of conference value by firstly using the exponential moving average (EMA) on Citation Ratio series $CR\_series_{A \to c_i}$ (or on Reference Ratio series $RR\_series_{A \to c_i}$) from each influenced conference from set $InfSet$ (the numerator in the line 14). Secondly, we calculate arithmetic average of values obtained from EMA (line 14). In this paper, we have chosen exponential moving average with its weights adding up to 1. The choice is motivated by the fact that in general case we believe that, especially in developing scientific areas, the ”fresh” articles should be more influential hence should be considered with higher weight. Importantly, thanks to the metric for Running Influence value, influence between two conferences in the same time period can be compared.

5 Validation

This section presents experiments conducted in order to validate the proposed model of Running Influence presented in Section 4.

Algorithm 3 Method for calculating overall Running Influence value for conference A

```
Input: Conference set $S \subset U$ for which we consider influence, conference $A \in S$ for which
    the Running Influence value is calculated, period $[t_s, t_e]$, maximal lag number $m$, significance
    level $\epsilon$, time window step $y$, $pf$ priority function
Output: Running Influence value $RI_A$ for conference $A$ during time period $[t_s, t_e]$
for all $c_i \in S \backslash \{A\}$ do
    for $t_i \Leftarrow t_s$ to $t_e$ with step $y$ do
        $CR\_series_{A \to c_i} += computeCR(A, c_i, t_i, pf)$
        $CR\_series_{c_i \to A} += computeCR(c_i, A, t_i, pf)$
    end for
    for $lag \Leftarrow 1$ to $m$ do
        $GCResult_{c_i \to A} += GCTest(CR\_series_{c_i \to A}, CR\_series_{A \to c_i}, lag)$
    end for
    $minGCResult_{c_i \to A} \Leftarrow min(GCResult_{c_i \to A})$
    if $minGCResult_{c_i \to A} < \epsilon$ then
        $InfSet += RI_{A \to c_i}$
    end if
end for
$RI_A = \frac{\sum_{j=1}^{|InfSet|} EMA(CR\_series_{A \to c_i})}{|Inf|}$
```

5.1 Data Description

In order to verify the proposed method, we have performed experiments in the real-world dataset consisting of the research papers: Microsoft Academic (MA) database [30]. The dataset contains 126 909 021 scientific publication entries and 528 682 289 citations between them [31], including the venue (conference) indication. This data was the base for creating the citation networks snapshots for each year, containing the number of citations from and to conferences. The summary of the dataset can be seen in Table 1.

Table 1. Statistics about used data from Microsoft Academic dataset

Parameter	**Number**
Number of Papers	1 349 526
Number of References	4 896 812
Number of selected conferences (AI/DM)	122
Total number of conferences in citation network	13 852

We limited all conferences and journals available in the dataset to a subset, as the subject of our experiments. In order to measure influence, we intended to choose a subset of already known and recognized conferences and journals. To achieve that, we have chosen the top-conferences/journals from Aminer list [32] using two groups: ”Artificial Intelligence and Pattern Recognition” (AI) and ”Databases and Data Mining” (DM). The Aminer rank uses H5-Index [24]. The final filtered list used in the experiments consists of 45 of Data Mining conferences and 77 Artificial Intelligence/Pattern Recognition conferences [3]. The experiments involving DM and AI conferences were performed independently from each other.

[3] A full list of all conferences and journals used in experiments is available at https://github.com/trzytematyczna/influence_csimq

The implementation and experiments were done using PostgreSQL[4] version 9.6 and R language[5] version 3.3.1 with the use of *lmtest* package [6] for Granger causality tests.

5.2 Experiments

The conducted experiments, based on the described dataset, consisted of determining and estimating the influence between conference groups using the time span between 1950 and 2015. Each Running Influence was calculated with respect to Definition 3. As it was previously mentioned, we took into consideration the values of the RI of each pair of conferences in order to calculate the final value of RI. The value of RI was calculated using Algorithm 3.

The examination of the model in terms of time, we have used three different priority functions: constant, linear and square root, shown in Figure 3. The choice of these functions was due to the fact that they emphasize different time-related aspects. A constant priority function emphasizes equally information in each time snapshot. On the other hand, both linear and square root functions perform annealing of historical data. In result, they put more emphasis on recent data, which is coherent with an intuitive approach of giving more priority to fresh papers.

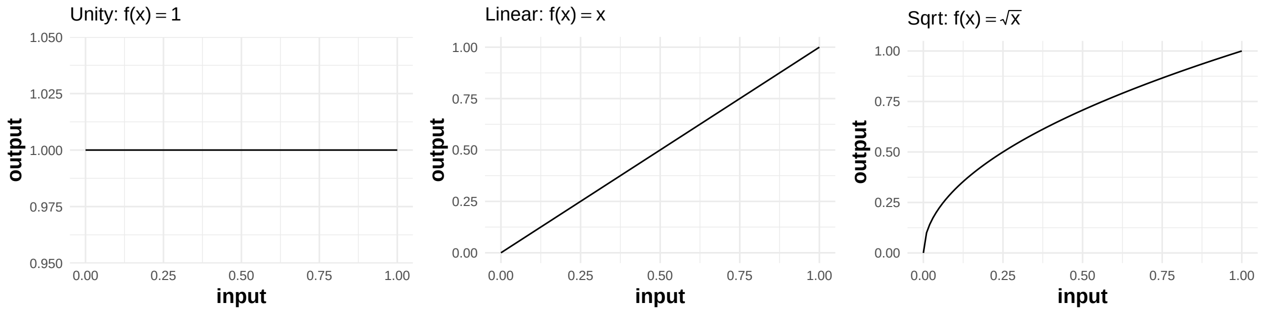
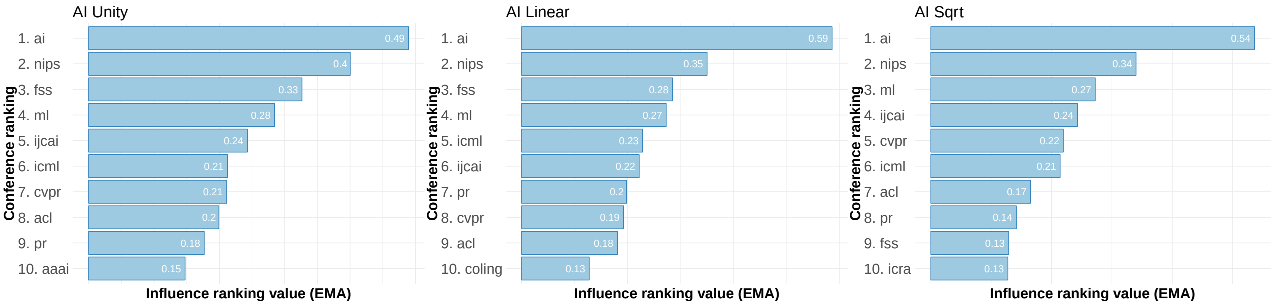


Figure 3. Three functions used as a priority function for calculating Citation Ratio series



(a) Citation Ratio and Unity priority function

(b) Citation Ratio and Linear priority function

(c) Citation Ratio and Sqrt priority function

Figure 4. The comparison of top 10 ranking of AI conferences – the result of RI value calculation (Algoritm 3), using Citation Ratio and three different priority functions

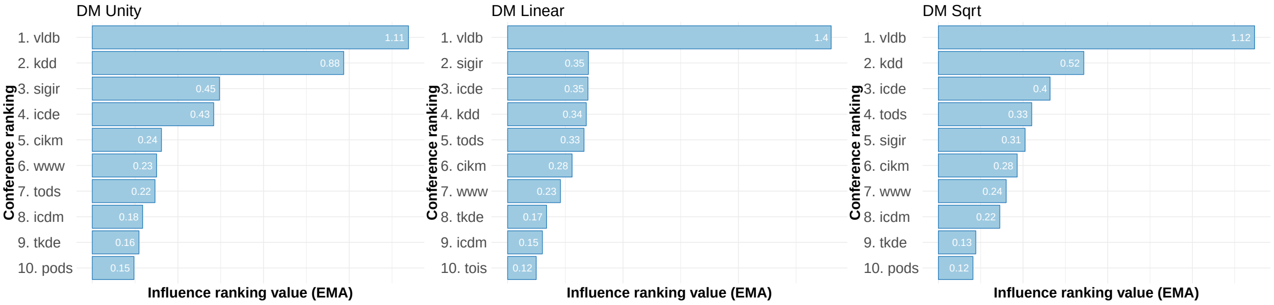
We performed several experiments. In order to investigate the time-dependency, we used the three priority functions on both DM and AI conferences utilizing Citation Ratio. The results for each of the functions using Citation Ratio for AI conferences in the form of top 10 influential conferences can be seen in Figure 4. Similarly, in case of DM conferences, the outcome utilizing Citation Ratio and three priority functions can be observed in Figure 5. On top of that, we also

[4] https://www.postgresql.org/

[5] https://www.r-project.org/

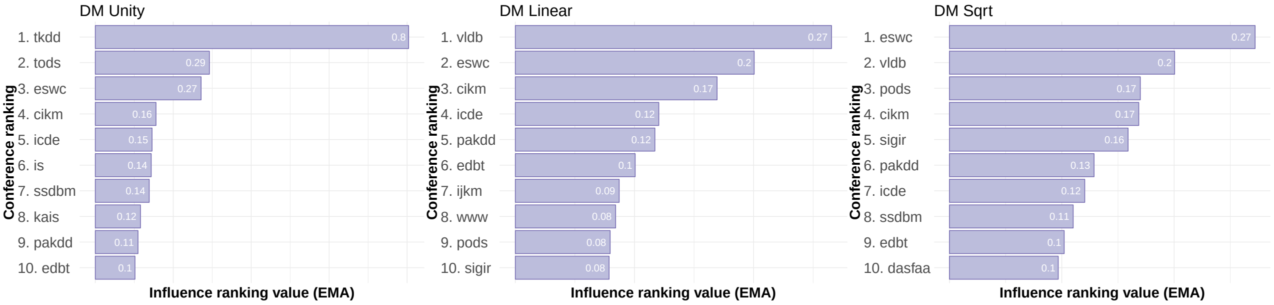
[6] https://cran.r-project.org/web/packages/lmtest/index.html

wanted to compare the methods of Citation Ratio and Reference Ratio, thus we also present the results of Reference Ratio with the use of three priority functions on the DM conferences in Figure 6.



(a) Citation Ratio and Unity priority function **(b)** Citation Ratio and Linear priority function **(c)** Citation Ratio and Sqrt priority function

Figure 5. The comparison of top 10 ranking of DM conferences – the result of RI value calculation (Algorithm 3), using Citation Ratio and three different priority functions



(a) Reference Ratio and Unity priority function **(b)** Reference Ratio and Linear priority function **(c)** Reference Ratio and Sqrt priority function

Figure 6. The comparison of top 10 ranking of DM conferences – the result of RI value calculation (Algorithm 3), using Reference Ratio and three different priority functions

In the case of ranks using Citation Ratio, the differences resulting from using three priority functions are not that drastic. For instance, in AI conferences the top two conferences - *ai* and *nips* are that broadly cited that no matter which priority function (that aims to stress the time importance) are used both of them are on top. This leads to the hypothesis that giving the difference stress on the time importance, using Citation Ratio it is possible to test whether a particular conference is truly influential, that is influential independently of time. If there is no difference between the results from different priority functions that emphasize (or not) first or last citation years, it means that the conference is indeed uniformly influential over the years. The changes of position of conferences in the rank depending on the priority function are due to the citation differences over the years. Unit priority function, does not emphasize the time in any way and treats all citations equally. In contrast, the approaches using linear and sqrt priority functions are more restrictive. This observation seems to confirm the hypothesis that considering time-dependency impacts the resulted influence dependencies.

Interesting to notice is also the vast differences between ranks for DM using Citation and Reference Ratios. Due to the fact that RR computed the ratio of citation between two particular conferences to all of the references to the cited conference the ranks include a different set of conferences. Along the conferences that are both in CR-based and RR-based ranks are conferences *www*, *cikm*, *vldb*, and *sigir*. It can be observed that in the case of RR results, the change of priority function results in significant change in the ranks. For instance, *tkdd*, having the 1st position in RR unity rank is not even included in the results from linear and sqrt priority functions.

Moreover, to show the differences between the results of using metrics Citation Ratio and Reference Ratio, we also created an influence dependency graph. Influence dependency graph is the result of calculating RI (see Algorithm 2) for each of the pairs of conferences, aggregated together, thanks to which we can observe all the influence relations between conferences. Figure 7 presents the influence dependency graph for DM conference set, obtained using Citation Ratio with unity priority function. In comparison, the influence dependency graph for DM with the usage of Reference Ratio and unity priority function can be seen in Figure 8. In each graph, the size of nodes symbolizes the hierarchy of the rank biggest node being most influential in the rank, smallest being least influential. The graphs additionally show the vast difference in the results from using the two metrics.

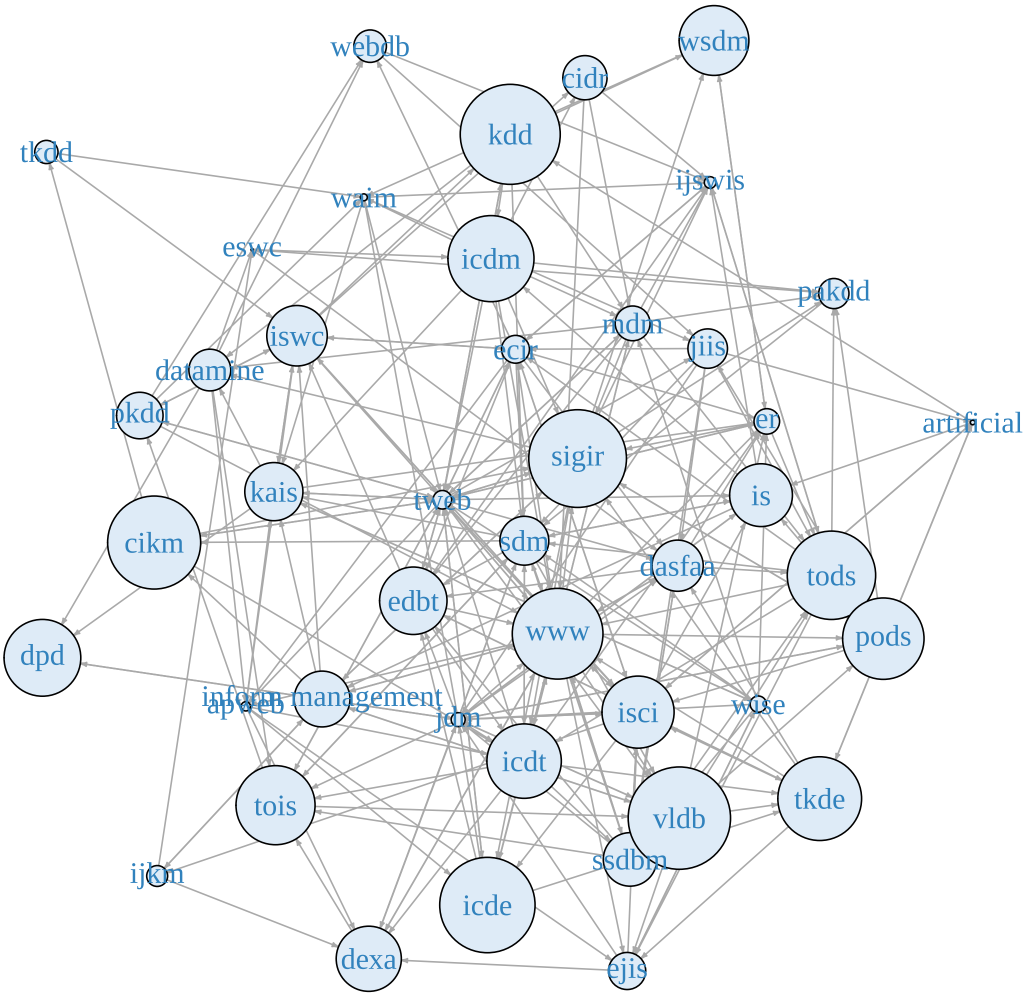


Figure 7. Influence dependency graph of DM conferences, using Citation Ratio and Unity priority function, where each edge represents pairwise Running Influence relation between conferences (Algorithm 2 using CR) for DM conference set

It is worth noticing that the influence value that is computed based on a number of citations does not guarantee the existence of a big number of meaningful connections (in terms of Granger causality). For instance, conference *tkdd*, which has the first place in rank created using Reference Ratio, can be seen in Figure 7, created using Citation Ratio, having almost none connections. Similarly, *vldb* being in top for CR-using rank and having big number of edges in CR-based graph, has a significantly smaller number of connections in case of the RR-based graph.

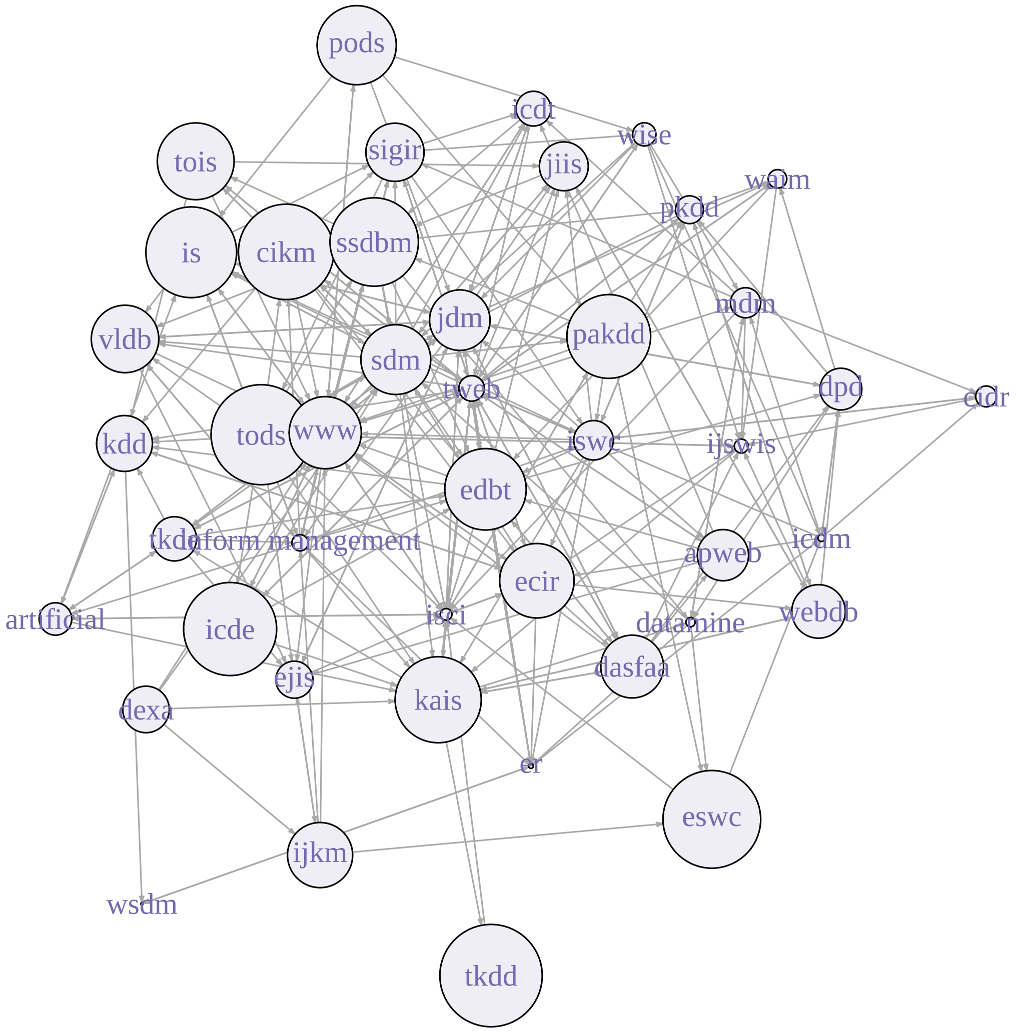
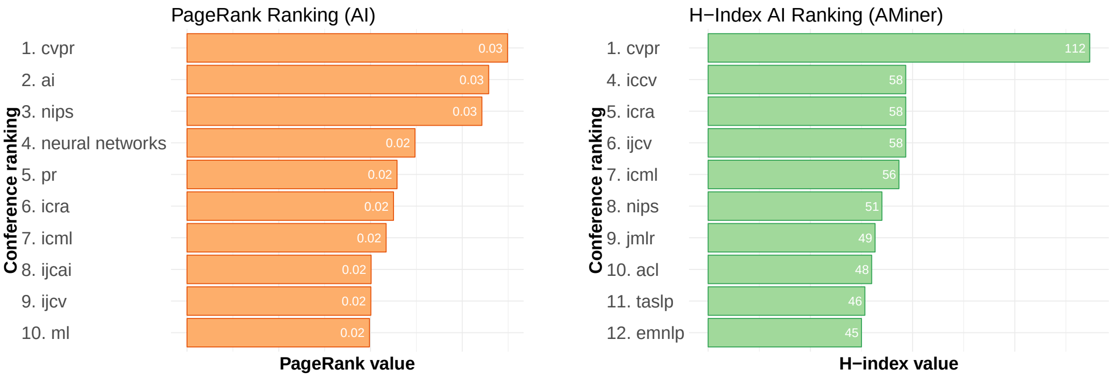


Figure 8. Influence dependency graph of DM conferences, using Reference Ratio and Unity priority function, where each edge represents pairwise Running Influence relation between conferences (Algorithm 2 using RR) for DM conference set

5.3 Comparison

For systems which deal with measuring relations within social networks, especially these considering influence, obtaining ground-trust knowledge is a widely known problem. Indeed, influence is actually a subjective notion, that originates in social sciences like sociology and psychology, making it ambiguous and hard to quantify. Depending on the particular system and research possibilities, each study tackles the problem of gathering the ground truth differently. In literature, in order to compare a proposed method with truth, we have observed the following trends: (1) presenting the analysis of the results containing only the proposed method [33], [28]; (2) comparing the proposed approach to real-world rankings [34]; (3) contrasting the presented method with another, similar-enough metrics; and (4) measuring ground-truth by utilizing or performing surveys or questionnaires [34], [35]. In our case, in order to provide both results and comparison of our experiments, we have chosen two, more versatile ways of comparison of the proposed method to other works: using real-world ranking and using a well-known influence metric – PageRank measure.

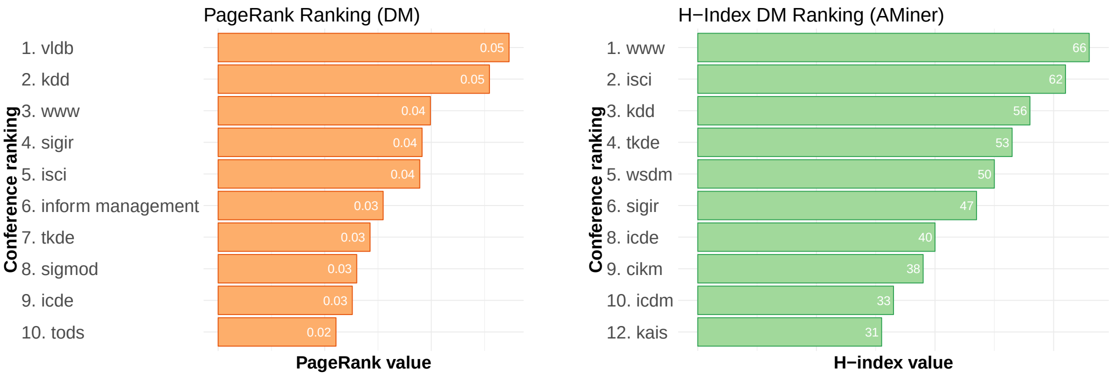
Real-world Conference Ranking For comparing results of our method with a real-world ranking, we have chosen the previously mentioned Aminer ranking. It consists of conference rankings, categorized by discipline, using H5-index. The rank and H5-score for AI conferences can be seen in Figure 9b, and for DM conferences in Figure 10b.



(a) Top 10 ranking of AI conferences using PageRank method

(b) Top 10 ranking of AI conferences using Aminer H-index rank

Figure 9. Ranks for AI conferences using PageRank (left) and Aminer H-index (right). Our database does not include some of the conferences present in the original Aminer H-index ranking, hence some positions in H-index ranks are missing.



(a) Top 10 ranking of DM conferences using PageRank method

(b) Top 10 ranking of DM conferences using Aminer H-index rank

Figure 10. Ranks for DM conferences using PageRank (left) and Aminer H-index (right). Our database does not include some of the conferences present in the original Aminer H-index ranking, hence some positions in H-index ranks are missing.

It can be seen that conferences such as *kdd*, *sigir*, or *cikm* are high in ranks both from our method and from H-index. This indicates that our approach is able to capture influence similar in a way to the assumed state-of-the-art method. However, while there are some similarities between the ranks of our method and H-index ranking, it is important to notice the differences between those approaches. Firstly, the H-index rank takes into consideration any citation from any journal whereas our method considers the subset of all conferences. Secondly, presented here H-index considers only the last 5 years of publication. While such information about latest trends of the conference is definitely useful, it differs from our method that takes a long-term view on citations of a conference. Moreover, due to the usage of priority functions such as linear or sqrt that put different stress on the 'historical' publications, the difference is even more visible, e.g. much lower places of *www* or *tkde* conferences.

Weighted PageRank In order to compare the results from the proposed model of Running Influence, we have chosen the weighted PageRank algorithm as a baseline due to two main reasons. Firstly, it is an accepted, well-established method, regarded as a *state-of-the-art* method for measuring influence [9]. Secondly, it is a well-known and popular approach, as it has been listed as one of ten most influential data-mining algorithms, particularly focused on prestige [36]. In particular, in order for PageRank to be as close to our approach as possible, we have chosen the weighted version of PageRank, where citations are considered as weights.

Having said that, it is important to notice that PageRank and proposed method of Running Influence differ in the way they define influence. PageRank assumes to calculate rank on a network with already specified, topologically-based influence. Moreover, it does not consider the time. In contrary, our method is using topological features of the network to further use it for calculating causal relations in the data. Furthermore, as mentioned, our method considers both time snapshots, and time-dependence of the citation networks.

In order to perform ranking with the use of weighted PageRank method, we have created a graph consisting of conferences as vertices and citations as edges. The lack of a notion of time for PageRank can be dealt with two-fold. Firstly, one could compute the PageRank values for each time snapshot. However, this leads to the problem of how to compute the overall value of PageRank using those snapshot PageRank values. Secondly, there is a possibility of aggregating citations from all used time snapshots in one graph, in order to use the citation numbers as weights for PageRank. Due to simplicity, we have chosen the aggregation method. Since in the experiments concerning DM and AI we use a Global version of both Citation and Reference Ratio for each conference, the PageRank graph contained the selected conferences, and also all conferences that ever cited or were cited by any of selected conferences. Networks for PageRank for each topic (DM/AI) were created separately, as disjoint graphs. The result in the form of top 10 conferences of using PageRank for AI can be seen in Figure 9a, whereas the outcome using DM conferences are shown in Figure 10a.

While PageRank is focused on topological properties of a network only, our method, apart from being topology-aware, is also adding the additional information about the quality of a connection – in particular – Citation Ratio modified by priority functions and ranks computed basing on the number of influenced entities and the volume of influence. One of the consequences of the fact that PageRank does not consider time and bases solely on network structure is that any newer conference that might be gaining influence, but has not long history, will not be included in rank from PageRank. In comparison, as our method includes time, a newer conference that is highly cited in ‘recent’ time can still get a high place in the rank. The instance of such situation can be seen in the case of younger conference *icdm*, that is not included in PageRank (Figure 10a), however, is included in all three ranks from CR-based method (Figure 5). Moreover, since H-index method takes into consideration last five years, we can say that it is not biased by the older conferences as PageRank is. Interestingly, we can observe that both H-index and our method based on CR has found some conferences not included in PageRank at all. In case of AI, conference *acl* is on 10th position (Figure 9b), while being on 7th, 8th and 9th position in the rank from our method (Figure 4). At the same time, *acl* is not included in PageRank. A similar situation occurs in the case of *cikm* conference (CR-based DM conferences, Figure 5), where it is placed on 5th and 6th position and 9th in H-index (Figure 10b), and not included in PageRank at all.

6 Conclusion

In this article, we have proposed a model for influence discovery and estimation for communities within citation networks that focuses on time-dependency aspect of influence. We presented two metrics, Citation Ratio and Reference Ratio for impact evaluation between communities, the latter being a novel proposition aiming to evaluate influence from a reference perspective. We conducted variety of the experiments using a real-world set of scientific conferences.

The results have shown that the model can identify and evaluate the influence between communities, also while comparing the obtained results to baseline methods of PageRank and

H-index. Furthermore, the conducted tests have shown that considering the time-dependency of publications in the model of influence basing on citations is important and not be omitted while dealing with community influence of conferences. Moreover, the presented outcome emphasized the differences between proposed two metrics for evaluating impact between communities.

As one of the directions for future work, further studies involving different communities are planned. In particular, the research comparing the results from experiments using communities created from author cliques versus conference communities could give very interesting insights on influence within scientific publication area. Moreover, a further extension of time emphasizing priority functions could also lead to improvement of the analysis of citation networks.

References

[1] M. Rakoczy, A. Bouzeghoub, K. Wegrzyn-Wolska, and A. Gancarski Lopes, *Users Views on Others – Analysis of Confused Relation-Based Terms in Social Network.* Cham: Springer International Publishing, 2016, pp. 155–174. [Online]. Available: https://doi.org/10.1007/978-3-319-48472-3_9

[2] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *arXiv preprint arXiv:1001.5285*, 2010.

[3] M. Kim, D. Newth, and P. Christen, "Modeling direct and indirect influence across heterogeneous social networks," in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM, 2013, pp. 1–9. [Online]. Available: https://doi.org/10.1145/2501025.2501030

[4] L. Liu, J. Tang, J. Han, and S. Yang, "Learning influence from heterogeneous social networks," *Data Mining and Knowledge Discovery*, vol. 25, no. 3, pp. 511–544, 2012. [Online]. Available: https://doi.org/10.1007/s10618-012-0252-3

[5] D. Kempe, J. M. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network." *Theory of Computing*, vol. 11, no. 4, pp. 105–147, 2015. [Online]. Available: https://doi.org/10.4086/toc.2015.v011a004

[6] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[7] P. Laflin, A. V. Mantzaris, F. Ainley, A. Otley, P. Grindrod, and D. J. Higham, "Discovering and validating influence in a dynamic online social network," *Social Network Analysis and Mining*, vol. 3, no. 4, pp. 1311–1323, 2013. [Online]. Available: https://doi.org/10.1007/s13278-013-0143-7

[8] M. E. Rakoczy, A. Bouzeghoub, A. L. Gancarski, and K. Wegrzyn-Wolska, "Influence in time-dependent citation networks," in *12th International Conference on Research Challenges in Information Science*, 2018, pp. 1–11. [Online]. Available: https://doi.org/10.1109/RCIS.2018.8406647

[9] R. Zafarani, M. A. Abbasi, and H. Liu, *Social media mining: an introduction*. Cambridge University Press, 2014. [Online]. Available: https://doi.org/10.1017/CBO9781139088510

[10] Y. Zhang, X. Li, and T.-W. Wang, "Identifying influencers in online social networks: The role of tie strength," *International Journal of Intelligent Information Technologies (IJIIT)*, vol. 9, no. 1, pp. 1–20, 2013. [Online]. Available: https://doi.org/10.4018/jiit.2013010101

[11] S. A. Myers, C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 33–41. [Online]. Available: https://doi.org/10.1145/2339530.2339540

[12] Y. Yang, N. V. Chawla, R. N. Lichtenwalter, and Y. Dong, "Influence activation model: A new perspective in social influence analysis and social network evolution," *arXiv preprint arXiv:1605.08410*, 2016.

[13] Y. Mehmood, N. Barbieri, F. Bonchi, and A. Ukkonen, “Csi: Community-level social influence analysis,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 48–63. [Online]. Available: https://doi.org/10.1007/978-3-642-40991-2_4

[14] P. Hui and S. Buchegger, “Groupthink and peer pressure: Social influence in online social network groups,” in *Social Network Analysis and Mining, 2009. ASONAM’09. International Conference on Advances in Social Network Analysis and Mining*. IEEE, 2009, pp. 53–59. [Online]. Available: https://doi.org/10.1109/ASONAM.2009.17

[15] W. Hu, Z. Gong, L. H. U, and J. Guo, “Identifying influential user communities on the social network,” *Enterprise Information Systems*, vol. 9, no. 7, pp. 709–724, 2015. [Online]. Available: https://doi.org/10.1080/17517575.2013.804586

[16] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, “Mining topic-level influence in heterogeneous networks,” in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 199–208. [Online]. Available: https://doi.org/10.1145/1871437.1871467

[17] B. Chikhaoui, M. Chiazzaro, and S. Wang, “Discovering and tracking influencer-influencee relationships between online communities,” in *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2015, pp. 1–9. [Online]. Available: https://doi.org/10.1109/DSAA.2015.7344846

[18] C. Tantipathananandh and T. Y. Berger-Wolf, “Finding communities in dynamic social networks,” in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 1236–1241. [Online]. Available: https://doi.org/10.1109/ICDM.2011.67

[19] N. Ohsaka, T. Akiba, Y. Yoshida, and K.-i. Kawarabayashi, “Dynamic influence analysis in evolving networks,” *Proceedings of the VLDB Endowment*, vol. 9, no. 12, pp. 1077–1088, 2016. [Online]. Available: https://doi.org/10.14778/2994509.2994525

[20] C. C. Aggarwal, S. Lin, and P. S. Yu, “On influential node discovery in dynamic social networks,” in *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, 2012, pp. 636–647. [Online]. Available: https://doi.org/10.1137/1.9781611972825.55

[21] M. Song, G. E. Heo, and S. Y. Kim, “Analyzing topic evolution in bioinformatics: investigation of dynamics of the field with conference data in dblp,” *Scientometrics*, vol. 101, no. 1, pp. 397–428, 2014. [Online]. Available: https://doi.org/10.1007/s11192-014-1246-2

[22] Y. Xie, Y. Sun, and L. Shen, “Predicating paper influence in academic network,” in *Computer Supported Cooperative Work in Design (CSCWD), 2016 IEEE 20th International Conference on*. IEEE, 2016, pp. 539–544. [Online]. Available: https://doi.org/10.1109/CSCWD.2016.7566047

[23] C. Moreira, P. Calado, and B. Martins, “Learning to rank academic experts in the dblp dataset,” *Expert Systems*, vol. 32, no. 4, pp. 477–493, 2015. [Online]. Available: https://doi.org/10.1111/exsy.12062

[24] J. E. Hirsch, “An index to quantify an individual’s scientific research output,” *Proceedings of the National academy of Sciences of the United States of America*, vol. 102, no. 46, p. 16569, 2005. [Online]. Available: https://doi.org/10.1073/pnas.0507655102

[25] B. Chikhaoui, M. Chiazzaro, and S. Wang, “A new granger causal model for influence evolution in dynamic social networks: The case of dblp,” in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI’15. AAAI Press, 2015, pp. 51–57.

[26] C. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969. [Online]. Available: https://doi.org/10.2307/1912791

[27] L. Dietz, S. Bickel, and T. Scheffer, “Unsupervised prediction of citation influences,” in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML ’07. New York, NY, USA: ACM, 2007, pp. 233–240. [Online]. Available: https://doi.org/10.1145/1273496.1273526

[28] B. Chikhaoui, M. Chiazzaro, S. Wang, and M. Sotir, "Detecting communities of authority and analyzing their influence in dynamic social networks," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 6, 2017. [Online]. Available: https://doi.org/10.1145/3070658

[29] J. Geweke, R. Meese, and W. Dent, "Comparing alternative tests of causality in temporal systems: Analytic results and experimental evidence," *Journal of Econometrics*, vol. 21, no. 2, pp. 161–194, 1983. [Online]. Available: https://doi.org/10.1016/0304-4076(83)90012-X

[30] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang, "Microsoft academic graph - 2016/02/05," available: http://academictorrents.com/details/1e0a00b9c606cf87c03e676f75929463c7756fb5, collected: 2016/02/05, downloaded: 06-2017.

[31] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15 Companion. New York, NY, USA: ACM, 2015, pp. 243–246. [Online]. Available: https://doi.org/10.1145/2740908.2742839

[32] AMiner, "Aminer computer science conference rank – artificial intelligence and pattern recognition," https://aminer.org/ranks/conf, accessed: 5-09-2017.

[33] J. H. Nguyen, B. Hu, S. Günnemann, and M. Ester, "Finding contexts of social influence in online social networks," in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM, 2013. [Online]. Available: https://doi.org/10.1145/2501025.2501028

[34] A. Rao, N. Spasojevic, Z. Li, and T. Dsouza, "Klout score: Measuring influence across multiple social networks," in *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE, pp. 2282–2289, 2015. [Online]. Available: https://doi.org/10.1109/BigData.2015.7364017

[35] G. Liu, F. Zhu, K. Zheng, A. Liu, Z. Li, L. Zhao, and X. Zhou, "Tosi: a trust-oriented social influence evaluation method in contextual social networks," *Neurocomputing*, vol. 210, pp. 130–140, 2016. [Online]. Available: https://doi.org/10.1016/j.neucom.2015.11.129

[36] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008. [Online]. Available: https://doi.org/10.1007/s10115-007-0114-2