A New Granger Causal Model for Influence Evolution in Dynamic Social Networks: The Case of DBLP

Belkacem Chikhaoui, Mauricio Chiazzaro, Shengrui Wang
Prospectus Laboratory, Department of Computer Science
University of Sherbrooke, Canada
{belkacem.chikhaoui, mauricio.chiazzaro, shengrui.wang}@usherbrooke.ca

Abstract

This paper addresses a new problem concerning the evolution of influence relationships between communities in dynamic social networks. A weighted temporal multigraph is employed to represent the dynamics of the social networks and analyze the influence relationships between communities over time. To ensure the interpretability of the knowledge discovered, evolution of the influence relationships is assessed by introducing the Granger causality. Through extensive experiments, we empirically demonstrate the suitability of our model for studying the evolution of influence between communities. Moreover, we empirically show how our model is able to accurately predict the influence of communities over time using random forest regression.

Introduction

Users in real-world social networks are organized into communities that are distinguished by number of users, preferences and interests, social influence, etc. (Zhang et al. 2013). Discovering and estimating social influence, with the aim of understanding how users affect each other, is an important research issue that has received considerable attention from the AI research community (Belák, Lam, and Hayes 2012; Mehmood et al. 2013; Zhang et al. 2013). In fact, detecting influential users in social networks and assessing their influence makes it possible to study information propagation in the network, which is very helpful in developing online advertisements, marketing campaigns and recommender systems (Barbieri and Bonchi 2014; Ye, Liu, and Lee 2012).

Much research has been conducted recently on detecting communities and studying influence in social networks (Chen, Wang, and Yang 2009; Leskovec, Lang, and Mahoney 2010). Detecting communities provides insight into the structure of the social network, whereas detecting influential users allows us to understand information dynamics and propagation and network evolution. Different types of influence have been proposed, such as pairwise influence (Goyal, Bonchi, and Lakshmanan 2010; Yin and Zhang 2012), social influence locality (Zhang et al. 2013), community influence (Belák, Lam, and Hayes 2012; Mehmood et al. 2013), topic influence (Liu et al. 2012; 2010; Tang et al. 2009), indirect influence (Kim, Newth, and Christen 2013; Shuai et al. 2012), and external influence (Myers, Zhu, and Leskovec 2012). However, these influence measures have been proposed for the purpose of detecting influential users and are not intended for assessing the influence between communities. Moreover, none of these measures studies the evolution of influence over time. With the rapid growth of social networks, detecting communities without assessing their overall influence in the network is now considered insufficient. This in turn suggests the need to study how communities influence each other and how this influence evolves over time. Finally, studying influence at the community level may reveal many more interesting patterns than merely looking at pairwise influence between users. These points constitute the major rationale for the work reported here.

In this paper, we propose an effective model for analyzing influence evolution in dynamic social networks. A weighted temporal multigraph is employed in order to represent the dynamics of social networks. The evolution of influence between communities is then assessed by incorporating the Granger causality (Granger 1969). Our model also makes it possible to predict the influence between communities using random forest regression. The combination of these methods yields an integrated framework for studying and predicting influence evolution. The major contributions of this paper can be summarized as follows:

1. Proposing an effective model for influence evolution using the Granger causality.

2. Combining weighted temporal multigraphs and the Granger causality for representing dynamic social networks and studying the evolution of inter-community influence.

3. Predicting the influence between communities over time using random forest regression.

The rest of the paper is organized as follows. First, we give an overview of related work in Section 2. Section 3 describes the proposed model in terms of weighted temporal multigraph representation, influence evolution and prediction. The results of our experiments on real social network datasets are presented in Section 4. Finally, Section 5 presents our conclusions.

Related work

In this section we discuss related work in the area of influence detection and assessment. We will focus mainly on influence between communities and its evolution over time.

Much research has been done in the social network field over the last decade. One of the most important areas of research in social networks concerns community detection and influential user extraction (Belák, Lam, and Hayes 2012; Liu et al. 2012; Mehmood et al. 2013; Chen, Wang, and Yang 2009; Zhang et al. 2013). Community detection allows one to pinpoint groups of users with common interests, tastes or goals (Dietz 2009), while influential user extraction allows one to locate those individuals that play a central role in the social network. Such users are those having high values of centrality or betweenness centrality (Kazuya, Wei, and Xiang-yang 2008).

Despite the considerable body of research on community detection, little work has been reported on assessing the influence between communities. Mehmood et al. (Mehmood et al. 2013) propose a community-level social influence measure for assessing the strength of influence between two different communities in directed social networks. Belák et al. (Belák, Lam, and Hayes 2012) propose a framework for cross-community influence analysis in discussion fora. The authors use the in-degree measure to assess the influence. Liu et al. (Liu et al. 2010) propose a topic-level influence in social networks. The authors also propose a method for calculating direct and indirect influence in social networks. Dietz et al. (Dietz, Bickel, and Scheffer 2007) propose an unsupervised prediction of citation influence in publication repositories. However, their model deals with the influence between papers and does not study the influence between communities.

Although existing models study influence between users or communities, none of them provides effective solutions to understand and predict the evolution of influence between communities in dynamic social networks. In this paper, we propose an effective model for studying the evolution of inter-community influence over time. Moreover, we build a method in our model to predict the influence between communities.

Our proposed model

In order to study the evolution of inter-community influence, our model needs a structure capable of representing the dynamics of the social network and displaying its state at each time instant. We resort to a multigraph formalism representation as a means to address this need and facilitate the assessment of influence evolution.

Weighted Temporal Multigraph

With the rapid evolution of social networks and their dynamics, basic graphs are unable to show the different aspects of the network dynamics. For this reason, we adapt the multigraph formalism in order to represent the dynamics of social networks. A multigraph is a graph in which multiple edges are permitted between two nodes. The rationale for using the multigraph representation is twofold: 1) it allows us to represent the temporal progression of the social network; and 2) it is a good visualization tool for the social network dynamics. In addition, in our model, a node represents a community as a whole and not an individual user, which is an extremely information-rich representation compared to classical graphs. As a result, our model deals with multigraphs of communities, which constitutes a new method of assessing influence evolution.

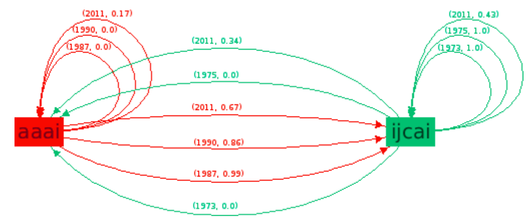


Figure 1: Example of a weighted temporal multigraph showing citation relationships between two communities, AAAI and IJCAI, in the DBLP dataset. Each edge here is marked by the time instant and its weight, the latter being calculated by Equation 1.

Given that influence between communities can be quantified and measured, we resort to the weighted temporal multigraph (WTMG) to represent the social network dynamics. A WTMG is a multigraph in which a weight (typically a real number) has been assigned to every edge at a time instant. Formally, we define a weighted temporal multigraph as $G = (V, E, T, W)$, where V is the set of nodes, $E \subseteq V \times T \times V \times W$ denotes a set of edges, T is a finite set of time instants, and W is a real-value function from E to the real numbers $\mathbb{R}^+$. A weighted temporal edge e in G is defined as an ordered quadruple $e = (u, v, t, w)$, where $u, v \in V$, with u possibly equal to v, are the origin and destination nodes, t is the time instant for the node u, and w is the weight of the edge e, which can be written as $w = W(e)$.

Figure 1 shows an example of a weighted multigraph, where the nodes represent two communities of the artificial intelligence discipline, the AAAI and IJCAI conferences.

As shown in Figure 1, the multigraph is a compact representation of graphs evolving over time. A node in a weighted temporal multigraph will have a matrix of influence values between itself and the other nodes at each time instant. The matrix of influence can be reduced to a vector of influence by accumulating and normalizing the influence values with respect to a time instant. As a consequence, each community $u \in G$ can be presented as a chronologically ordered series of influence vectors over time. To illustrate this point, without loss of generality, let the AAAI and IJCAI conferences be two artificial intelligence communities in the graph G of the DBLP dataset. Let the function $W()$ be the number of citations of the papers of one conference by the papers of the other conference for each year. If no citation is reported between the two conferences at a particular time instant, then the edge weight will simply be zero. The Figure 2 illustrates citation relationships between the AAAI and IJCAI communities at different time instants in the DBLP and Arnet Miner

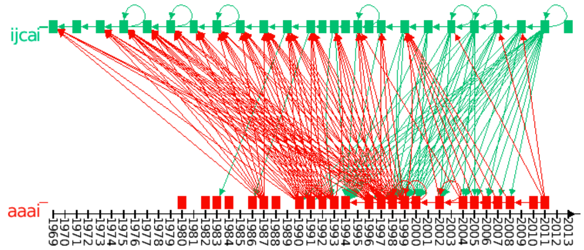


Figure 2: Citation relationships between the two communities AAAI and IJCAI in the DBLP and Arnet Miner datasets.

datasets.

As shown in Figure 2, each pair of nodes will have two chronologically ordered series of influence values over time. For example, at each time instant t_i, a community $u \in G$ will have a certain number of citations $\{N(j \rightarrow u, t_i) | j \in v\}$ of another community $v \in G$. Each community is composed of a certain number of papers. Therefore, a normalized citation weight is computed to represent the citation influence (CI) (Dietz, Bickel, and Scheffer 2007) between the two communities at a particular time instant t_i. Note that the citation influences $u \rightarrow v$ and $v \rightarrow u$ are different and should both be computed using the following formula:

$$CI(u \rightarrow v, t_i) = \frac{\sum_{j \in v} N(j \rightarrow u, t_i)}{M_{v,t_i}}, \tag{1}$$

where $N(j \rightarrow u, t_i)$ represents the number of citations of community u by a paper $j \in v$ at time instant t_i, and M_{v,t_i} represents the number of all citations made by community v from the other communities in graph G at time instant t_i. The citation influence values obtained at each time instant will be used to study the influence evolution, as described in the next section. Table 1 shows the citation influence values computed between the AAAI and IJCAI communities using Equation 1 at different time instants. The labels in Table 1 reflect the year in which an article from one community was published, assuming that article cited the other community articles published in any year (current or previous).

Table 1: $CI(AAAI \rightarrow IJCAI)$ and $CI(IJCAI \rightarrow AAAI)$ values computed using Equation 1.

	1991	**1993**	**1997**	**1999**	**2005**	**2011**
AAAI→IJCAI	0.854	0.947	0.698	0.401	0.055	0.666
IJCAI→AAAI	0.032	0.142	0.416	0.787	0.471	0.335

As shown in Table 1, the value (0.854) of the citation influence AAAI → IJCAI obtained during 1991 can be interpreted as indicating that 85.4 % of the citations made by the IJCAI community are from the AAAI community. Similarly, only 3.2 % of the citations made by the AAAI community are from the IJCAI community.

Influence Evolution Analysis

Influence between communities is a challenging research issue and little work has been reported on detecting the influence between communities. Moreover, studying the evolution of influence in dynamic social networks is a new problem. We propose an effective method for studying the evolution of influence between communities. Our method is based on the use of the Granger causality (Granger 1969) to infer influence evolution.

The rationale for incorporating the Granger causality in assessing influence evolution is twofold. First, it helps develop a more effective model for discovering hidden knowledge from the data. Secondly, it makes the discovered knowledge more interpretable from both the statistical and the semantic standpoints. The Granger causality has gained tremendous success across many domains due to its simplicity, robustness, extendability and it involves no hypothesis about the data (Ioannis, David, and Wan 2000). The Granger causality was initially developed for analyzing the effect of one time series on another. Formally, suppose we have two stationary time series $CI(u \rightarrow v) = \{CI(u \rightarrow v)(t)_{t \in T}\}$ and $CI(v \rightarrow u) = \{CI(v \rightarrow u)(t)_{t \in T}\}$ and we intend to study whether one influences (Granger causes) the other or not. The regression formulation of Granger causality states that $CI(u \rightarrow v)$ influences (Granger causes) $CI(v \rightarrow u)$ if the past values of $CI(u \rightarrow v)$ are helpful in predicting the future values of $CI(v \rightarrow u)$. If there is no influence between the two time series, the null hypothesis holds. The two regression formulas are presented below:

$$H_0 : CI(v \rightarrow u)(t) = \sum_{l=1}^{L} a_l CI(v \rightarrow u)(t - l) + \epsilon_1 \tag{2}$$

$$H_1 : CI(v \rightarrow u)(t) = \sum_{l=1}^{L} a_l CI(v \rightarrow u)(t - l) + \sum_{l=1}^{L} b_l CI(u \rightarrow v)(t - l) + \epsilon_2, \tag{3}$$

where L is the maximal time lag, a_l and b_l are the regression variable coefficients, and ϵ_1 and ϵ_2 are the residual terms, which are independent and identically distributed according to a standard Gaussian $N(0; \sigma^2)$. If Equation 3 is a significantly better model than Equation 2, we conclude that time series $CI(u \rightarrow v)$ Granger causes time series $CI(v \rightarrow u)$. Among other techniques, the models above can be tested using the Granger Sargent test (Granger 1969), defined as follows:

$$F = \frac{(RSS_{\epsilon_1} - RSS_{\epsilon_2})/L}{(RSS_{\epsilon_2})/(n - 2L)} \sim F(L, n - 2L), \tag{4}$$

where RSS_{ϵ_1} is the "restricted" residual sum of squares under H_0 , RSS_{ϵ_2} is the "unrestricted" residual sum of squares under H_1, and n is the number of observations. The Granger causality is assessed for the two time series $CI(u \rightarrow v)$ and $CI(v \rightarrow u)$ in both directions in order to discover whether the influence between them is bi-directional or unidirectional. Thus, by using the Granger causality principle, we will be able to assess the evolution of the influence between communities at each time instant.

Influence Prediction

Predicting the influence between communities over time involves evaluating the influence that one community might exert on the other at a specific time instant. For instance, what is the predicted value of the influence that community AAAI might exert on community IJCAI during the year 2015? Computing the predicted influence value is challenging. We use a random-forest-based regression method to predict the influence between two communities. Random-forest-based regression has been proved to be effective in terms of prediction accuracy (Khan 2014). Our method utilizes the weighted temporal multigraph state at time t in order to predict the state at $t+1$. It proceeds by first learning the evolution of inter-community influence values and then predicting the future influence between communities using the learned model.

Formally, a random forest is a predictor consisting of a collection of tree-structured predictors $\{h(\mathbf{x}, \Theta_k), k = 1, 2, ..., K\}$, where $\mathbf{x}$ represents the observed input vector of length p with associated random vector $\mathbf{X}$ and Θ_k are independent and identically distributed (iid) random vectors. Each tree $h(\mathbf{x}, \Theta_k)$ is constructed employing a different bootstrap sample of the training dataset, using the algorithm in (Breiman 2001). As we mentioned above, since our goal is to predict the influence between two communities, we focus on the regression aspect for which we have a numerical output, Y (Segal 2004).

A random forest for regression is an unweighted average over the collection $\bar{h}(\mathbf{x}) = (1/K)\sum_{k=1}^{K} h(\mathbf{x}, \Theta_k)$. According to (Segal 2004), as $k \to \infty$, the law of large numbers ensures that the following holds:

$$E_{\mathbf{X},Y}(Y - \bar{h}(\mathbf{X}))^2 \to E_{\mathbf{X},Y}(Y - E_\Theta h(\mathbf{X}, \Theta))^2 \quad (5)$$

The quantity $E_{\mathbf{X},Y}(Y - E_\Theta h(\mathbf{X}, \Theta))^2$ is the prediction error of the random forest, designated as PE_f^*. The average prediction error for an individual tree $h(\mathbf{X}, \Theta)$ can be computed as follows:

$$PE_{tree}^* = E_\Theta E_{\mathbf{X},Y}(Y - h(\mathbf{X}, \Theta))^2 \quad (6)$$

Validation

This section presents the datasets used and discusses the results obtained for the evolution and prediction of influence between communities.

Dataset

To validate our proposed model, we used the DBLP dataset, which contains information about articles, authors, conferences, and dates. However, it does not provide information about citation relationships between papers. We therefore augmented the DBLP dataset with paper citation information by incorporating the Arnet Miner citation network dataset. We merged the two datasets based on the article title to form one complete, rich network dataset. Table 2 shows the details of each dataset.

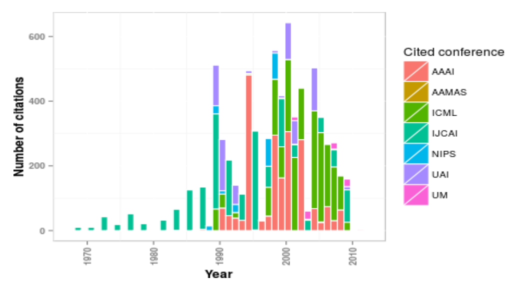
In our work, we selected well-known research disciplines such as Artificial Intelligence (AI), Data Mining (DM) and Human Computer Interaction (HCI) to validate our proposed model. In each discipline, top-ranked conferences, according to the Microsoft conference ranking system, have been selected as the communities for which influence evolution and prediction are evaluated.

Table 2: Datasets used for validation

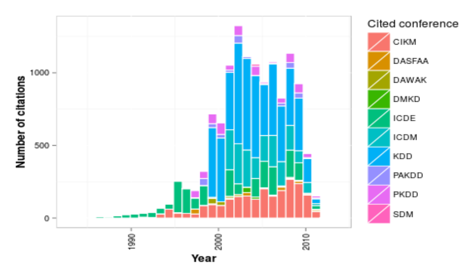
	Number of papers	Citations	Date
DBLP	2 712 770	No citations	08-08-2014
Arnet Miner	2 244 021	4 354 534	25-05-2014

Influence Evolution

Influence evolution is assessed between each pair of conferences in each discipline. Before assessing the influence evolution using the Granger causality, we computed all of the citation influences between all conferences at each time instant using formula 1. The values computed for citation influence between communities provide the basis on which the evolution of that influence is assessed using the Granger causality. Figure 3 on the next page shows a graphical representation of the distribution of citations between all communities in the AI and DM disciplines.



(a) AI



(b) DM

Figure 3: Citation values of communities at each time instant for each discipline.

To assess the influence evolution, we performed two types of validation: 1) a validation using all the citation influence values computed for all time instants, and 2) a validation using citation influence values computed for intervals between time instants to track the influence evolution.

Influence Evolution Using All Time Instants In this validation, we used the citation influence values computed between communities for all time instants. The influence evolution is assessed between each pair of communities. For example, to assess the influence evolution between communities AAAI and IJCAI, we computed the citation influence values using formula 1 for the AAAI community by taking the AAAI $\overset{Cites}{\rightarrow}$ IJCAI citations for all time instants (i.e., from 1969 to 2013). Similarly, we computed the citation influence values for the IJCAI community by taking the IJCAI $\overset{Cites}{\rightarrow}$ AAAI citations for all time instants.

Once the citation influence vectors have been computed, the Granger causality can be assessed. We computed the F-statistic and P-values in order to assess the Granger causality. Then, based on these results, the direction of influence between each pair of communities is established. Tables 3, 4 and 5 show the influence relationships obtained between each pair of communities for each discipline.

Table 3: Influence relationships obtained for the AI discipline.

	F-value	P-value	Influence direction
AAAI$\overset{Cites}{\rightarrow}$ICML	0.1729	0.6800	
ICML$\overset{Cites}{\rightarrow}$AAAI	**5.4735**	**0.0251**	**AAAI$\longmapsto$ICML**
AAAI$\overset{Cites}{\rightarrow}$IJCAI	**7.0334**	**0.0010**	**IJCAI$\longmapsto$AAAI**
IJCAI$\overset{Cites}{\rightarrow}$AAAI	0.0511	0.9843	
ICML$\overset{Cites}{\rightarrow}$UAI	**4.0914**	**0.0154**	**UAI$\longmapsto$ICML**
UAI$\overset{Cites}{\rightarrow}$ICML	3.1663	0.0392	
UAI$\overset{Cites}{\rightarrow}$IJCAI	**5.4709**	**0.0251**	**IJCAI$\longmapsto$UAI**
IJCAI$\overset{Cites}{\rightarrow}$UAI	4.1302	0.0497	
UM$\overset{Cites}{\rightarrow}$IJCAI	**49.4017**	**1.6462e-10**	**IJCAI$\longmapsto$UM**
IJCAI$\overset{Cites}{\rightarrow}$UM	2.4947	9.8434e-02	

Table 4: Influence relationships obtained for the DM discipline.

	F-value	P-value	Influence direction
CIKM$\overset{Cites}{\rightarrow}$ DAWAK	**18.4613**	**0.0001**	**DAWAK$\longmapsto$ CIKM**
DAWAK$\overset{Cites}{\rightarrow}$ CIKM	0.2210	0.8046	
CIKM$\overset{Cites}{\rightarrow}$ ICDE	**5.4344**	**0.0177**	**ICDE$\longmapsto$ CIKM**
ICDE$\overset{Cites}{\rightarrow}$ CIKM	0.7044	0.5708	
DMKD$\overset{Cites}{\rightarrow}$ ICDE	**51.6178**	**2.1726e-06**	**ICDE$\longmapsto$ DMKD**
ICDE$\overset{Cites}{\rightarrow}$ DMKD	0.0455	8.3364e-01	
DMKD$\overset{Cites}{\rightarrow}$ KDD	**6.5576**	**0.0209**	**KDD$\longmapsto$ DMKD**
KDD$\overset{Cites}{\rightarrow}$ DMKD	1.7726	0.2017	
DMKD$\overset{Cites}{\rightarrow}$ PKDD	**11.1662**	**0.0015**	**PKDD$\longmapsto$ DMKD**
PKDD$\overset{Cites}{\rightarrow}$ DMKD	0.1078	0.8985	
KDD$\overset{Cites}{\rightarrow}$ DASFAA	**3.9801**	**0.0448**	**DASFAA$\longmapsto$ KDD**
DASFAA$\overset{Cites}{\rightarrow}$ KDD	0.6789	0.5242	
PAKDD$\overset{Cites}{\rightarrow}$ CIKM	**5.3874**	**0.0182**	**CIKM$\longmapsto$ PAKDD**
CIKM$\overset{Cites}{\rightarrow}$ PAKDD	1.7249	0.2247	
PAKDD$\overset{Cites}{\rightarrow}$ ICDE	**7.8417**	**0.0058**	**ICDE$\longmapsto$PAKDD**
ICDE$\overset{Cites}{\rightarrow}$ PAKDD	0.0764	0.9268	
PAKDD$\overset{Cites}{\rightarrow}$ PKDD	**19.4611**	**0.0004**	**PKDD$\longmapsto$ PAKDD**
PKDD$\overset{Cites}{\rightarrow}$ PAKDD	0.0227	0.8819	
SDM$\overset{Cites}{\rightarrow}$ KDD	**81.9723**	**2.4255e-07**	**KDD$\longmapsto$ SDM**
KDD$\overset{Cites}{\rightarrow}$ SDM	0.1208	9.4570e-01	

As shown in tables 3, 4 and 5, our model is able to determine which community of a pair of communities is more influential. For example, in the AI discipline, the results show that the AAAI community significantly influences the ICML community, based on the F-values (5.4735 $>$ 0.1729) and P-values (0.0251 $<$ 0.6800) computed for ICML$\overset{Cites}{\rightarrow}$AAAI and AAAI$\overset{Cites}{\rightarrow}$ICML respectively. Therefore, the AAAI community can be considered as an influential community in the AI discipline. Consequently, one of the potentials of our model is the ability to detect influential communities using the Granger causality tests. To this end, we propose the following definition for an influential community:

Table 5: Influence relationships obtained for the HCI discipline.

	F-value	P-value	Influence direction
CHI$\overset{Cites}{\rightarrow}$ HCI	0.2581	0.6160	
HCI$\overset{Cites}{\rightarrow}$ CHI	**7.3180**	**0.0123**	**CHI$\longmapsto$ HCI**
CHI$\overset{Cites}{\rightarrow}$ HUC	1.4453	0.2410	
HUC$\overset{Cites}{\rightarrow}$ CHI	**5.3463**	**0.0296**	**CHI$\longmapsto$ HUC**
CSCW$\overset{Cites}{\rightarrow}$ IUI	1.4379	0.2421	
IUI$\overset{Cites}{\rightarrow}$ CSCW	**9.3921**	**0.0053**	**CSCW$\longmapsto$ IUI**
CSCW$\overset{Cites}{\rightarrow}$ UIST	1.1675	0.2906	
UIST$\overset{Cites}{\rightarrow}$ CSCW	**7.2489**	**0.0127**	**CSCW$\longmapsto$ UIST**
ECSCW$\overset{Cites}{\rightarrow}$ HUC	**4.0247**	**0.0331**	**HUC$\longmapsto$ ECSCW**
HUC$\overset{Cites}{\rightarrow}$ ECSCW	0.2218	0.8028	
HCI$\overset{Cites}{\rightarrow}$ HUC	0.1673	0.6860	
HUC$\overset{Cites}{\rightarrow}$ HCI	**13.8550**	**0.0010**	**HCI$\longmapsto$ HUC**
HUC$\overset{Cites}{\rightarrow}$ IUI	1.5109	0.2437	
IUI$\overset{Cites}{\rightarrow}$ HUC	**5.6946**	**0.0105**	**HUC$\longmapsto$ IUI**
HUC$\overset{Cites}{\rightarrow}$ UIST	**5.9564**	**0.0089**	**UIST$\longmapsto$ HUC**
UIST$\overset{Cites}{\rightarrow}$ HUC	0.2873	0.7531	
ISWC$\overset{Cites}{\rightarrow}$ IUI	**4.9097**	**0.0178**	**IUI$\longmapsto$ ISWC**
IUI$\overset{Cites}{\rightarrow}$ ISWC	0.2254	0.8000	

Definition 1 *Let $CI(u \rightarrow v)$ and $CI(v \rightarrow u)$ be two citation influence vectors. Community u is influential if $CI(u \rightarrow v)$ Granger causes $CI(v \rightarrow u)$, and the P-value $\leq \epsilon$*

For example, if $\epsilon = 0.03$, the communities IJCAI, AAAI and UAI can be considered as influential communities in the AI discipline. The same observations can be generalized for the DM and HCI disciplines. We choose $\epsilon = 0.03$ to indicate that the results are highly significant. The communities KDD, ICDE, DAWAK, PKDD and CIKM are thus the influential communities in the DM discipline. Similarly, the communities CHI, HCI, CSCW, HUC, UIST and IUI are the influential communities in the HCI discipline.

Influence Evolution Using Time Intervals The purpose of this validation is to show how the influence between communities evolves over time. To this end, we computed the Granger causality between each pair of communities for different time intervals. For example, for the AI discipline, we computed the Granger causality for the intervals 1969 to 2009 and 1969 to 2010 (the influence results for the time interval 1969 to 2011 are presented in Table 3). The results obtained allow us to understand how the influence between communities evolves from the first time interval to the last. Tables 6 and 7 show the results obtained for the influence between each pair of communities in the AI discipline for each time interval.

Table 6: Influence results obtained for the AI discipline for the time interval 1969 to 2009.

	F-value	P-value	Influence direction
AAAI $\stackrel{Cites}{\rightarrow}$ ICML	0.2402	0.6272	
ICML $\stackrel{Cites}{\rightarrow}$ AAAI	**4.5101**	**0.0412**	**AAAI$\longmapsto$ICML**
AAAI $\stackrel{Cites}{\rightarrow}$ IJCAI	**7.9633**	**0.0016**	**IJCAI$\longmapsto$AAAI**
IJCAI $\stackrel{Cites}{\rightarrow}$ AAAI	0.0711	0.9315	
ICML $\stackrel{Cites}{\rightarrow}$ UAI	0.2567	6.1576e-01	
UAI $\stackrel{Cites}{\rightarrow}$ ICML	**29.7501**	**4.8242e-06**	**ICML$\longmapsto$UAI**
UAI $\stackrel{Cites}{\rightarrow}$ IJCAI	**5.9411**	**0.0203**	**IJCAI$\longmapsto$UAI**
IJCAI $\stackrel{Cites}{\rightarrow}$ UAI	4.3208	0.0454	
UM $\stackrel{Cites}{\rightarrow}$ IJCAI	**57.9152**	**5.0014e-11**	**IJCAI$\longmapsto$UM**
IJCAI $\stackrel{Cites}{\rightarrow}$ UM	5.0955	1.2441e-02	
AAAI $\stackrel{Cites}{\rightarrow}$ UAI	1.3887	0.2675	
UAI $\stackrel{Cites}{\rightarrow}$ AAAI	**3.4245**	**0.0312**	**AAAI$\longmapsto$UAI**

Table 7: Influence results obtained for the AI discipline for the time interval 1969 to 2010.

	F-value	P-value	Influence direction
AAAI $\stackrel{Cites}{\rightarrow}$ ICML	0.3087	0.5821	
ICML $\stackrel{Cites}{\rightarrow}$ AAAI	**5.0843**	**0.0306**	**AAAI$\longmapsto$ICML**
AAAI $\stackrel{Cites}{\rightarrow}$ IJCAI	**8.2204**	**0.0013**	**IJCAI$\longmapsto$AAAI**
IJCAI $\stackrel{Cites}{\rightarrow}$ AAAI	0.0453	0.9557	
ICML $\stackrel{Cites}{\rightarrow}$ UAI	**23.4571**	**6.2522e-07**	**UAI$\longmapsto$ICML**
UAI $\stackrel{Cites}{\rightarrow}$ ICML	5.3754	9.9034e-03	
UAI $\stackrel{Cites}{\rightarrow}$ IJCAI	**5.4526**	**0.0255**	**IJCAI$\longmapsto$UAI**
IJCAI $\stackrel{Cites}{\rightarrow}$ UAI	3.9706	0.0543	
UM $\stackrel{Cites}{\rightarrow}$ IJCAI	**59.9145**	**2.2366e-11**	**IJCAI$\longmapsto$UM**
IJCAI $\stackrel{Cites}{\rightarrow}$ UM	5.2113	1.1191e-02	

The results reported in Tables 6 and 7 clearly show how the influence values change between communities at each time interval. For example, for ICML$\stackrel{Cites}{\rightarrow}$AAAI, the F-value increased from 4.5101 in 2009, to 5.0843 in 2010, to 5.4735 in 2011. Similarly, the P-value decreased from 0.0412 in 2009, to 0.0306 in 2010, to 0.0251 in 2011. This means that the community AAAI is gaining influence for the ICML community. An important observation can be made for UAI$\stackrel{Cites}{\rightarrow}$ICML. Indeed, the F-value decreased from 29.7501 in 2009 to 5.3754 in 2010. Similarly, the P-value increased from $4.8242e{-}06$ in 2009 to $9.9034e{-}03$ in 2010. This variation indicates a change in the direction of the influence between the two communities: i.e., in 2009, **ICML $\longmapsto$ UAI**, while in 2010, **UAI $\longmapsto$ ICML**. The influence direction **ICML $\longmapsto$ UAI** established in 2009 has been re-established again in 2011, as shown in Table 3. This example clearly demonstrates that our model is able to study and analyze the evolution of influence between communities. Table 6 also shows that the AAAI community influences the UAI community (**AAAI$\longmapsto$UAI**) in 2009. However, this influence relationship no longer holds in 2010 (F-value = 0.5400, P-value = 0.4674) and 2011 (F-value = 0.3067, P-value = 0.5831).

Influence Prediction

An influence prediction is calculated for each pair of communities, and the results are averaged for each discipline. We used ten-fold cross-validation for training and test. We computed the Correlation Coefficient (CC), the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) in order to measure the accuracy of our model. For comparison purposes, we used two other regression models, the linear regression model and the multilayer perceptron model, to highlight the suitability and performance of our random forest regression model compared with the two other models. The rational of using these methods is that the linear regression is considered as the reference regression model, and the multilayer perceptron is the most commonly used model for comparing regression methods given its performance and reliability. Table 8 shows the results obtained for the three models.

Table 8: Influence prediction results by discipline using RF(Random Forest), LR(Linear Regression) and MP(Multilayer Perceptron).

	AI			DM			HCI		
	CC	MAE	RMSE	CC	MAE	RMSE	CC	MAE	RMSE
RF	**0.811**	**0.068**	**0.138**	**0.876**	**0.042**	**0.090**	**0.950**	**0.024**	**0.062**
LR	0.618	0.107	0.188	0.816	0.055	0.107	0.927	0.0317	0.075
MP	0.712	0.127	0.180	0.799	0.061	0.116	0.906	0.051	0.089

As shown in Table 8, our model achieves the highest coefficient of correlation of the three models. Moreover, our model results in lower prediction error than the others. This demonstrates the efficiency of our model and its suitability for predicting the influence between communities.

Conclusion

In this paper we have investigated a new problem concerning the evolution of influence between communities in dynamic social networks. We have proposed a new model, based on the Granger causality, for studying influence evolution. Our model utilizes a weighted temporal multigraph to represent the dynamics of the social network. The evolution of inter-community influence was studied by incorporating the Granger causality and utilizing the influence values computed between communities at each time instant. We have also proposed a method based on random forest regression to predict the influence between communities.

We have illustrated the effectiveness and suitability of our model through extensive experiments on the DBLP dataset. The experimental results demonstrate that our model is able to study influence evolution over time and that it can accurately predict the influence between communities by minimizing the prediction error.

It will be interesting in the future to conduct more experiments using other dynamic social networks such as Facebook, Youtube and Twitter, and study the evolution of the influence between online communities.

References

Barbieri, N., and Bonchi, F. 2014. Influence maximization with viral product design. In *SDM*.

Belák, V.; Lam, S.; and Hayes, C. 2012. Cross-community influence in discussion fora. In *ICWSM*.

Breiman, L. 2001. Random forests. *Mach. Learn.* 45(1):5–32.

Chen, W.; Wang, Y.; and Yang, S. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, 199–208. New York, NY, USA: ACM.

Dietz, L.; Bickel, S.; and Scheffer, T. 2007. Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, 233–240. New York, NY, USA: ACM.

Dietz, L. 2009. Modeling shared tastes in online communities. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*.

Goyal, A.; Bonchi, F.; and Lakshmanan, L. V. 2010. Learning influence probabilities in social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, 241–250. New York, NY, USA: ACM.

Granger, C. W. J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37(3):424–438.

Ioannis, A.; David, A.; and Wan, M. M. 2000. Non-linear granger causality in the currency futures returns. *Economics Letters* 68(1):25 – 30.

Kazuya, O.; Wei, C.; and Xiang-yang, L. 2008. Ranking of closeness centrality for large-scale social networks. In *Frontiers in Algorithmics*, 186–195.

Khan, I. 2014. Bias-corrected quantile regression forests for high-dimensional data. In *Proceedings of the 13th International Conference on Machine Learning and Cybernetics*. IEEE.

Kim, M.; Newth, D.; and Christen, P. 2013. Modeling direct and indirect influence across heterogeneous social networks. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, SNAKDD '13, 9:1–9:9. New York, NY, USA: ACM.

Leskovec, J.; Lang, K. J.; and Mahoney, M. 2010. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, 631–640. New York, NY, USA: ACM.

Liu, L.; Tang, J.; Han, J.; Jiang, M.; and Yang, S. 2010. Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, 199–208. New York, NY, USA: ACM.

Liu, L.; Tang, J.; Han, J.; and Yang, S. 2012. Learning influence from heterogeneous social networks. *Data Mining and Knowledge Discovery* 25(3):511–544.

Mehmood, Y.; Barbieri, N.; Bonchi, F.; and Ukkonen, A. 2013. Csi: Community-level social influence analysis. In *ECML/PKDD (2)*, 48–63.

Myers, S. A.; Zhu, C.; and Leskovec, J. 2012. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, 33–41. New York, NY, USA: ACM.

Segal, M. R. 2004. Machine learning benchmarks and random forest regression. Technical report, Center for Bioinformatics & Molecular Biostatistics. UC San Francisco.

Shuai, X.; Ying, D.; Jerome, B.; Shanshan, C.; Yuyin, S.; and Jie, T. 2012. Modeling indirect influence on twitter. *IJSWIS* 8(4):20–36.

Tang, J.; Sun, J.; Wang, C.; and Yang, Z. 2009. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, 807–816. New York, NY, USA: ACM.

Ye, M.; Liu, X.; and Lee, W.-C. 2012. Exploring social influence for recommendation: A generative model approach. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, 671–680. New York, NY, USA: ACM.

Yin, Z., and Zhang, Y. 2012. Measuring pair-wise social influence in microblog. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 502–507.

Zhang, J.; Liu, B.; Tang, J.; Chen, T.; and Li, J. 2013. Social influence locality for modeling retweeting behaviors. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI'13, 2761–2767. AAAI Press.