Data Min Knowl Disc (2015) 29:1152–1177
DOI 10.1007/s10618-015-0402-5

Dynamic inference of social roles in information cascades

Sarvenaz Choobdar · Pedro Ribeiro · Srinivasan Parthasarathy · Fernando Silva

Received: 2 March 2014 / Accepted: 12 January 2015 / Published online: 5 February 2015

Abstract Nodes in complex networks inherently represent different kinds of functional or organizational *roles*. In the dynamic process of an information cascade, users play different roles in spreading the information: some act as seeds to initiate the process, some limit the propagation and others are in-between. Understanding the roles of users is crucial in modeling the cascades. Previous research mainly focuses on modeling users behavior based upon the dynamic exchange of information with neighbors. We argue however that the structural patterns in the neighborhood of nodes may already contain enough information to infer users' roles, independently from the information flow in itself. To approach this possibility, we examine how network characteristics of users affect their actions in the cascade. We also advocate that temporal information is very important. With this in mind, we propose an unsupervised methodology based on ensemble clustering to classify users into their social roles in a network, using not only their current topological positions, but also considering their history over time. Our experiments on two social networks, Flickr and Digg,

Responsible editors: Toon Calders, Floriana Esposito, Eyke Hüllermeier and Rosa Meo.

S. Choobdar (✉) · P. Ribeiro · F. Silva
CRACS and INESC-TEC, Faculdade de Ciencias, Universidade do Porto,
R. Campo Alegre, 1021, 4169-007 Porto, Portugal
e-mail: sarvenaz@dcc.fc.up.pt

P. Ribeiro
e-mail: pribeiro@dcc.fc.up.pt

F. Silva
e-mail: fds@dcc.fc.up.pt

S. Parthasarathy
The Ohio State University, Columbus, OH, USA
e-mail: srini@cse.ohio-state.edu

show that topological metrics indeed possess discriminatory power and that different structural patterns correspond to different parts in the process. We observe that user commitment in the neighborhood affects considerably the influence score of users. In addition, we discover that the cohesion of neighborhood is important in the blocking behavior of users. With this we can construct topological fingerprints that can help us in identifying social roles, based solely on structural social ties, and independently from nodes activity and how information flows.

Keywords Structural role mining · Information cascade · Social role · Ensemble clustering · Complex networks

1 Introduction

Users in online social networks get engaged in different social activities such as sharing and exchanging information. Information propagation models study how an idea gets spread in social networks. These studies mainly consider users activity and their neighbors activity to model the process. In an information cascade, users behave differently: some are more active in terms of adopting new ideas, some cause blockage and others are more influential in spreading the ideas. Understanding social behavior of users is important in modeling the information propagation in many diverse phenomena, including adoption of new ideas, spread of infectious diseases, computer virus epidemics on the Internet, viral marketing campaigns, and information cascades in online social networks (Adamic and Adar 2005; Easley and Kleinberg 2010; Myers et al. 2012; Wang et al. 2012).

Users behavior is essentially modeled based on the history of their activity and their friends' activity, that is, on the information flow in itself. The structural connectivity of the network has comparatively received much less attention. Regardless, it has been shown that network characteristics of users also affect their activity. For instance, Leman et al. show how users' influence is correlated to centrality measures (Ghosh and Lerman 2012).

In this work we investigate precisely how the social status of users relates to their structural position in the network. Nodes at different topological positions, such as centers of stars, members of cliques and peripheral nodes may have different functions. The roles are defined using structural measurements of the node and its neighborhood. More specifically, we study information propagation of *stories* in social networks, and we concentrate on the effects of structural patterns on two different properties: level of *influence* and *blockage rate*. We categorize users into different roles in a social activity from these two points of view. User influence is related to the cascade size a user can cause, that is, the amount of other users that receive stories propagated by such cascade. Blockage rate amounts to the number of stories a user does not repost, normalized to the total number of received stories. We use network characteristics of users to classify them into social groups and try to find a correspondence between topological positions and social role.

Our end goal is to use pure structural properties to reveal social activity and to discover the essential connectivity principles behind social activities. For this we

propose a novel dynamic framework to classify nodes that is able to incorporate time information. We ensure that current node roles are close to previous roles when the connectivity does not deviate much, and that a significant change on the structure of the network also implies an updated set of roles reflecting that new topology. Our methodology includes a new evolutionary classification method that incorporates ensemble clustering (Strehl and Ghosh 2003), by which we combine multiple partitionings of a set of objects without accessing the original features. Ensemble clustering has been shown to be a robust and accurate methodology (Topchy et al. 2004; Fred and Jain 2002; Gionis et al. 2005; Strehl and Ghosh 2003). Nevertheless, it has essentially been used for static data where different partitionings of the same dataset are given. In a recent paper by Lancichinetti and Fortunato, the dynamic communities in a network are explored by cluster aggregation(Lancichinetti and Fortunato 2012). To the best of our knowledge, ensemble clustering has not been used for evolutionary clustering. Here, for the first time, we use this concept in temporal data to extract the grouping of data regarding their feature set and their history. We use a temporal weighting function in the aggregation of users' behavior over time.

Our main research contributions in this work are the following:

- An in-depth analysis of how pure topological features are related to the roles of users in information cascades, namely their influence and blockage rate. We explore the effect of several structural properties on the social activity and show that there is a correlation between the role of a user and its position in the network and that these metrics have the ability to distinguish among different roles;
- A new unsupervised evolutionary clustering methodology, capable of categorizing users in information cascades. We propose a novel incorporation of ensemble clustering and a framework that is able to discover social roles that both reflect the structure of the network at present time and is consistent with past roles. We compare our performance against other baseline methods, showing that we can outperform them, and we use our methodology to infer social activity roles.

The remainder of the article is organized as follows. Section 2 reviews previous approaches and related problems. Section 3 examines the relations between structural properties of users and their social activity in an information cascade in two datasets of Digg and Flickr. Section 4 describes the proposed ensemble clustering method for extracting dynamic social roles of users. Section 5 provides results about the experimental evaluation of our proposed method on social networks. Section 6 gives the final comments on the obtained results and concludes the article.

2 Related work

Information propagation in social networks has been widely studied for a number of years from different aspects, we group them into two categories. The first category includes research works that study the process of influence spread and how the information propagates from one to another. The second category includes research studies that focus on characterizing users in order to find a set of nodes with maximal influence.

In the first category, several influence models have been proposed and studied, and the most popular ones are the linear threshold model (LT) and the independent cascade model (IC), by Kempe et al. (2003). These models study spread of influence through social networks, where the influence probabilities between users are predefined. Saito et al. (2008) predict the influence probabilities in independent cascade models of propagation using maximum likelihood estimation, and Goyal et al. (2010) study the probabilities in the threshold model by counting the number of correlated social actions. They both consider the temporal nature of influence of users.

In the second category, users' influence is measured using structural models of influence such as PageRank and in-degree centrality in the network (Kwak et al. 2010), number of followers, mentions, retweets (Lee et al. 2010; Cha et al. 2010), or the size of the information cascades (Bakshy et al. 2011). Earlier studies of social influence and propagation showed that the most influential bloggers were not necessarily the most active (Agarwal et al. 2008). Temporal information has been used in modeling influence using the influence-passivity score (Romero et al. 2011). An important aspect of information dissemination is the study of parameters that stop the contagion. Steeg et al. showed that many cascades grow slower than expected and do not reach "epidemic" proportions (Ver Steeg et al. 2011). Their study on Digg data showed that multiple exposures to the same information does not affect the probability of voting. The same phenomena is seen on Flickr data where the photos are not spread in a quick and viral fashion throughout the social network (Cha et al. 2009). Although the structure of the Flickr social network holds small-world properties, which in theory says a piece of information will spread quickly and widely through social links, photos on Flickr are spread with delay (Cha et al. 2012). This study concludes that propagation is not only due to activity of users but also due to information availability at the time of users' activity.

Using multiple sources of information may raise the complexity of the analysis, but also brings more resolution to the problem. Tang et al. (2009) leverages another source of information for finding topic-specific influence. They use topic distribution of users in conjunction with a social network of users to build a factor graph model, and propose a topical affinity propagation on the factor graph to automatically identify the topic-specific social influence. Zhou and Liu (2013) integrated three sources of information to derive the influence group of users. They defined a new similarity matrix between users based on three sources of information including a social network of users, activity networks and influence networks. They proposed a clustering algorithm based on k-means that divides users into homogeneous groups regarding the derived similarity matrix. They combined social influence based similarity between each pair of users by unifying the self-similarity and multiple co-influence similarity scores through a weight function with an iterative update method.

All referred papers use both the activity log of users and their social network to characterize the influence process. However, in this paper we only use topological properties of users to categorize their role in the influence spread.

Role extraction is an exploratory task where no *a priori* class for nodes is available and the role assignment of nodes is desirable. This is different from label acquisition which is commonly defined in the literature as determining the label for a node in a network that is partially labeled. Normally, for node labeling it is assumed that at least

some of the nodes have a predefined label and only the labels for remaining nodes are predicted using relational classifiers (Taskar et al. 2002; Gallagher and Eliassi-Rad 2010). For a static network, role extraction is defined as the process of finding groups of nodes with similar properties. In other words, this is a clustering task where nodes are grouped, not based on their connectivity, but because they hold a similar position in the network. This has been studied by other researchers, where nodes with the most outstanding properties are detected as singular motifs using outliers detection methods (Costa et al. 2009). Henderson et al. (2012) found roles of nodes regarding their properties in their neighborhood by non-negative matrix factorization. In their method, the matrix of node-role is derived from matrix factorization of node-features and features-role matrices, and the number of roles is determined by a Minimum Description Length (MDL) method (Rissanen 1978).

In our previous works (Choobdar et al. 2011, 2012), a two-phase general methodology was designed to characterize time evolving networks. In the first step of this methodology, nodes are grouped by k-means clustering and classified based on their role in the network. In the second step a method is proposed to study the evolution of the network using a supervised approach. In this method a set of events happening in the network is defined for the roles in the network. We then find the predefined events happening in the network and the rules that describe them by using association rule mining. Rossi et al. (2012), used the methodology proposed by Henderson et al. (2012) for dynamic role extraction. They measure a set of features for nodes at each time snapshot. Then, by stacking all the node-by-feature matrices, they derive the matrix of feature-roles by factorizing the stacked node-by-feature matrix and iteratively generate the matrix of node-role for each time. Role discovery methods are essentially unsupervised. However a more supervised approach for role discovery is presented by Zhao et al. (2013) where they used structural properties of users to infer their pre-defined social statuses of users. They proposed a probabilistic model to integrate users' social properties and network features for prediction of users roles. Danilevsky et al. (2013) studied role discovery in hierarchical topical communities. They defined the role of a user as his contribution to the community. In their method contextual properties of users is models and structural properties are not concerned.

3 Relation between network topology and social activity

In this section we analyze the role of users in information cascades and how network characteristics of individual nodes affect their social activities. We quantitatively study the relation between a number of structural properties of users in a network and two aspects of information cascades: user influence and user blockage.

3.1 Network data

Throughout this entire section we will be using two different datasets, coming from two well known and established internet communities: *Digg*[1] and *Flickr*.[2] Both include

[1] http://www.isi.edu/~lerman/downloads/digg2009.html.

[2] http://socialnetworks.mpi-sws.org/datasets.html.

Table 1 Summary of datasets

Data	#Users	#Objects (story/photo)	#Links	Network time interval	Time granularity
Digg	71,367	3,553	1,731,658	5 years	Three months
Flickr	914,400	4,000	18,595,048	2 years	One month

a static social network including social relationships between users and a dynamic evolving network describing information propagation.

Digg is a news aggregator in which users can submit links to interesting news stories and they can rate these stories by voting on them. Users also can designate other users as friends. More specifically, each user has a list of followers (fans who follow him) and a list of followees (friends whom he follows). All activities are visible to his fans, including all stories he submitted or voted for. We use the Digg data collected by Lerman and Ghosh (2012) which contains the friendship network of users and all the posts submitted during one month, including the id, submitter id, voters for each post and the date of votes. This dataset includes 3,018,197 votes on 3,553 popular stories made by 139,409 users and the social network of active users (who have at least one vote) containing 71,367 users and 1,731,658 friendship links. We built our social network from active users and their connections, where active users are those who voted for at least one story.

Flickr is a popular photo and video hosting website with a large community of users. We use data collected by Cha et al. (2009), which includes a social friendship network of users and information propagation from one user to another. The associated mechanism is similar to Digg, but instead of URLs, photos are shared and voted. This dataset contains data of 104 days (starting Nov 2, 2006) on 34,734,22 favorite markings of 11,267,320 photos. The social network has 1,620,392 users and 33,140,018 edges. We sample 4000 photos from those whose number of favorite marking is higher than 100. The social network includes all users who have marked the selected photos as favorites and all their connections in the original data.

Table 1 summarizes the statistics of the data we used. Fig. 1a shows the cumulative distribution of degree, eigenvector centrality and clustering coefficient of users. In both datasets, the degree distribution follows a scale-free distribution as it is common to degree distributions in real-world complex networks.

3.2 User influence

In order to study how topology of networks relates to influence level of users in an information cascade we measure the influence score of users for information propagation in networks. We assign social roles to users using the influence score in an information cascade. There are two different definitions for the empirical influence of users: (1) size of cascade initiated by a user (Bakshy et al. 2011); (2) number of votes a user's stories receives from his fans (Ghosh and Lerman 2010). These definitions are limited to submitters but in a cascade other users also play important roles in spreading the information and have some levels of influence in the cascade. Hence, we adopt the second definition for all users as influence score:

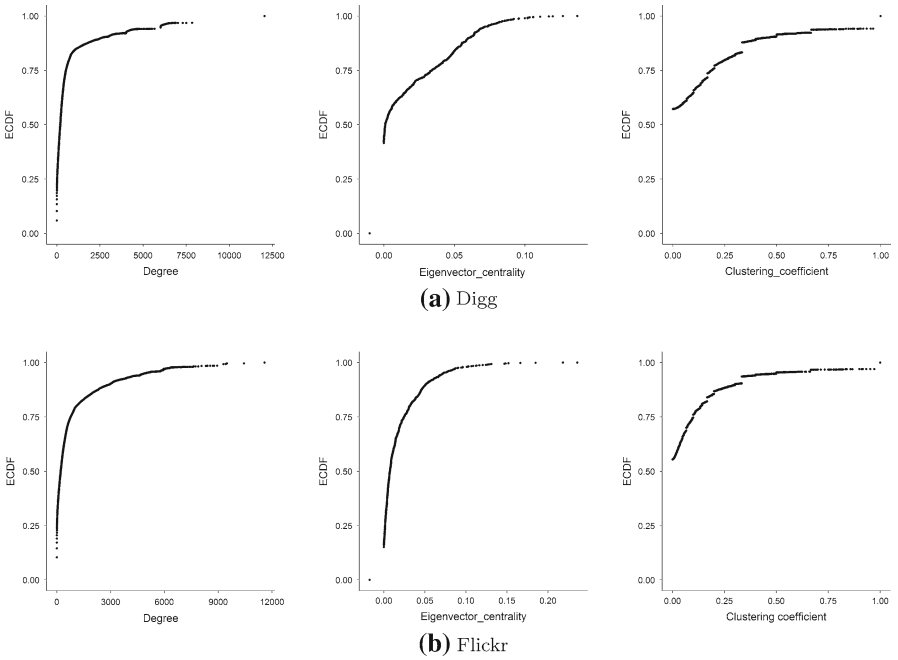


Fig. 1 The empirical cumulative distribution function (ECDF) of structural properties of social networks in Digg and Flickr data

$$Influence(u_i) = \sum_j \frac{votes_f(s_j)}{|posts(u_i)|}, s_j \in posts(u_i) \quad (1)$$

where $votes_f(s_j)$ is the number of votes story s_j receives from fans of user u_i after user u_i has voted, and $posts(u_i)$ is the set of all stories submitted or voted by user u_i.

For example, in Digg data when a user submits a story it becomes visible to his fans. Some of his fans may like the story and vote for it, making the story visible to their fans as well, and this process goes on. All users are important in spreading the information but at different levels. Figure 2 shows the distribution of influence scores for users in Digg and Flickr social networks. We categorize users into different groups regarding their influence score. We use equal width discretization to factorize the influence value and classify users into five groups from non-influential, to highly influential. In Fig. 2 groups are highlighted with different colors. The influence models mentioned in Sect. 2 are basically built on the individual users features and do not take into account the neighborhood properties. In this paper we study the effect of neighborhood structure on users' influence. We examine the correlation of structural features on the influence of users regarding reachability and commitments of users.

Reachability of users is important for spreading information and most of the influence models are based on this property. We quantify reachability of a user in a network by using "degree centrality" defined as the number of users directly connected to a

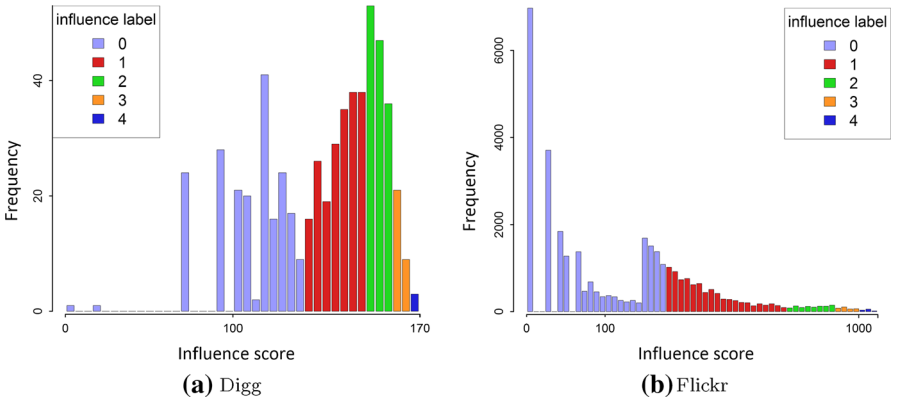
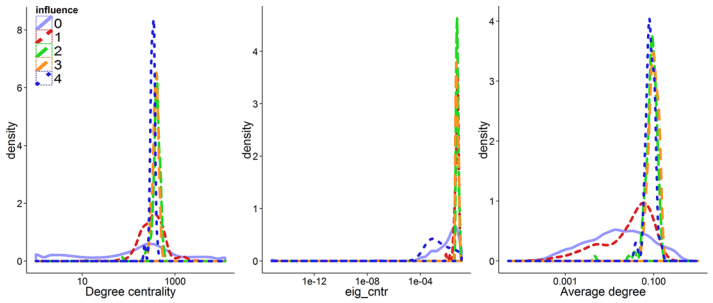


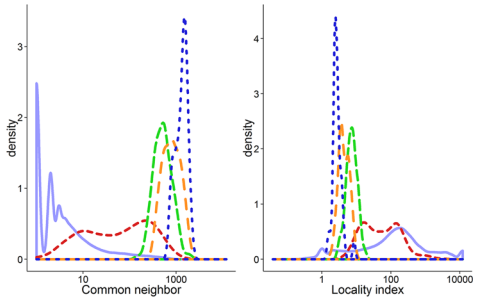
Fig. 2 The histogram of influence score of users in Digg and Flickr data (Color figure online)

user. We also study the degree distribution in the neighborhood of a user by measuring the "average degree in neighborhood". This property represents the "2-hop" reach of individuals in the network. Out of three centrality measures of betweenness, closeness, and eigenvector, we have selected eigenvector which had higher distinguishing power. We examine the eigenvector centrality (Bonacich 2007) of users, which rank users regarding their importance in the network. This centrality metric acts similarly to degree centrality. However, it gives higher score to the nodes which are themselves connected to high score nodes. In other words, the quality of neighbors of a node is accounted in eigenvector centrality. Figure 3 shows the distinguishing power of these three reachability measurements, we can see that the distribution on all five influence groups for eigenvector centrality have highest distinctions and then average degree in neighborhood is more distinguishing than degree centrality.

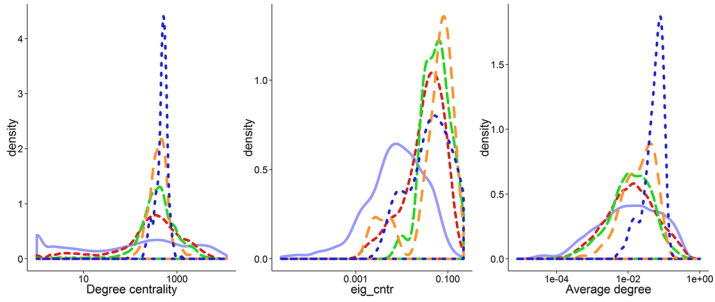
Commitment of users to their neighborhood is another feature that we study. Commitment shows how well a user is connected to his neighborhood comparing to the whole network (Granovetter 1985). We study the effect of this feature on influence of users by measuring two structural properties: "locality index" (Loc) and "common neighbors" (CN). Loc is the ratio of the number of connections between neighbors and the rest of the network to the number of connections between neighbors of a user. CN is the number of common neighbors between a neighbor of a user and his neighbors' connections. Figure 3 illustrates the distribution of commitments properties (Loc, CN) for influential groups for users in Digg and Flickr social network. We can see a peak in Loc disribution graph in Fig. 3. This peak belongs to the fourth group of influence category which includes users with very high influence score. These users all have very low locality index and varying in short range, which means they are located in a dense neighborhood and their neighbors are more connected to themselves than to the rest of network. The plot of common neighbor in Fig. 3 confirms this results, as we see that users in this group share many neighbors. We observe that the probability distribution of Loc distinctly changes from a influence group to another. Generally, we can see that commitment properties can better distinguish users at different influence level comparing to reachability properties.



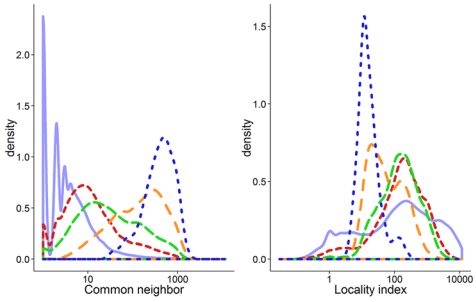
(a) Reachability features for Digg



(b) Commitment features for Digg



(c) Reachability features for Flickr



(d) Commitment features for Flickr

Fig. 3 Correlation between social roles of users in an information cascade network and their network characteristics including degree centrality, average degree, eigenvector centrality, common neighbor (CN) and locality index (Loc) at different levels of influence in Digg and Flickr social networks (Color figure online)

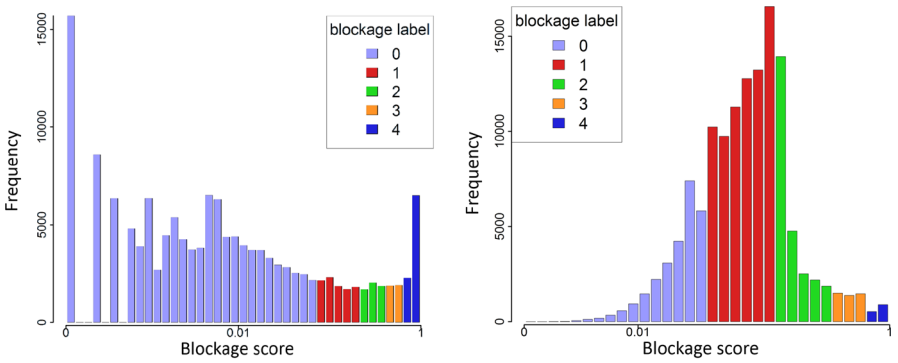


Fig. 4 The histogram of blockage rate of users in Digg and Flickr data (Color figure online)

3.3 Cascade blockage

Many of the cascades grow far slower than expected from their initial spread and fail to reach epidemic proportions (Lerman et al. 2012). The network structure somehow limits the growth of cascades. In this section we study the role of users in stopping a cascade. We define the blockage rate as the probability of not voting for a story if at least one of the users' friends has voted for it. We estimate this probability as the fraction of stories visible to a user and those that he did not vote for. We can formulate it as:

$$Blockage(u_i) = \frac{\sum b(s_j)}{|received(u_i)|}, s_j \in received(u_i) \quad (2)$$

where $received(u_i)$ is the set of stories visible to user u_i and $b(s_j)$ is 0 if u_i repost s_j and 1 otherwise.

We categorize users into different groups based on the blockage rate using an equal width discretization method. Figure 4 shows the distribution of blockage rate of users in Digg and Flickr data. Based on this distribution we have five groups of users, shown in different colors from non-blockers to blockers. Non-blockers are users with very low blockage score shown in Fig. 4. We investigate the correlation of blockage rate of users against three structural properties of users in a network.

Triadic closure is an important property that represents the triangular structure of a network (Granovetter 1973). The local triangular structures in networks is a fundamental feature that causes spread of information in networks (Iribarren and Moro 2009). To incorporate it in our method we use the local clustering coefficient of each user (Kossinets and Watts 2006; Guo et al. 2011). This measures the number of triangles (cliques of size 3) a user i is involved in, normalized by the number of triplets of connected nodes (not necessarily cliques) that include the same user i. The clustering structure of networks causes multiple exposures to a story and this may limit the spread of information (Lerman et al. 2012). Figure 5 shows the distribution of clustering coefficient of users at different levels of blockage rate in two social networks of Digg and Flickr.

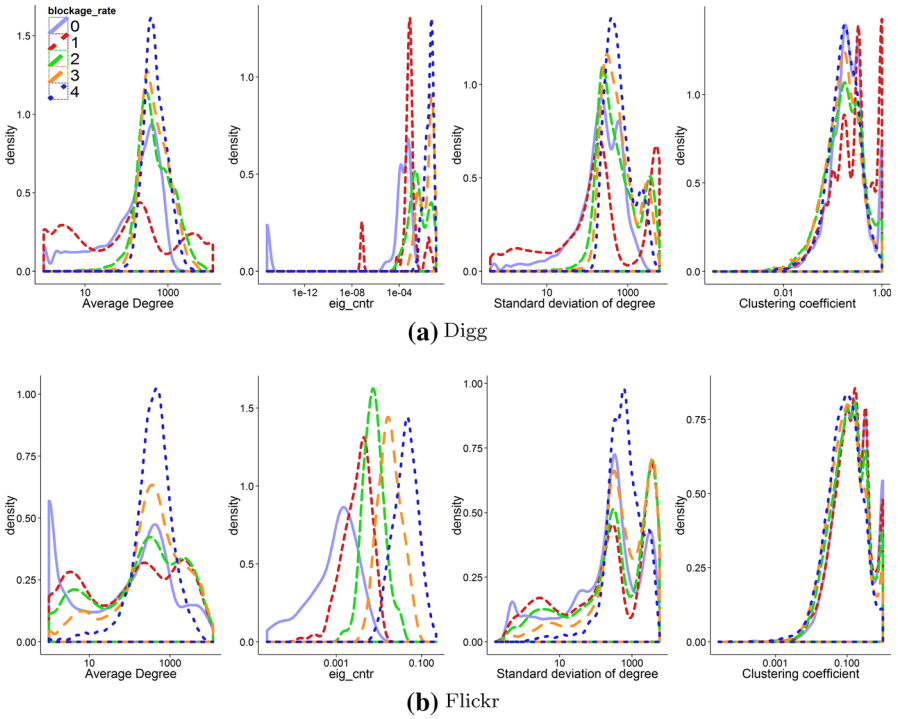


Fig. 5 Correlation between social roles of users in an information cascade and their network characteristics including average degree, eigenvector centrality, standard deviation of degree in neighborhood and clustering coefficient in Digg and Flickr social networks (Color figure online)

Besides the triadic closure we examine the relation of blockage rate with other network characteristics. We assess the effect of neighborhood cohesion on the behavior of users. To quantitatively measure the cohesion in the neighborhood of a user, we use standard deviation and euclidean mean of degree of connected users (Fig. 5). The cohesion of the neighborhood of a user appears to be more effective on determining the user's voting behavior. As we can see from Fig. 5, distributions of average degree and standard deviation of degree for users with high blockage rate are more shifted to the right and are centered around higher average comparing to the groups with lower blockage rate.

We also examine how centrality of users affects blocking behavior of users. We use eigenvector centrality of users and one can see that eigenvector centrality of users at different blockage rates has different ranges. This is more obvious in the Flickr dataset, as we can see in Fig. 5.

In summary, network characteristics of users affect their social activities and can be used to categorize users into different social roles. The correlation analysis among cascade properties (influence score and blockage score) and structural properties of users are depicted in Table 2. As we can see the correlation values are not strong nor negligible. We note that most of the structural properties are weakly correlated and independently can not infer social roles of users. In the following sections we

Table 2 Correlation between structural properties of users and their social activity in Digg network

Influence score	Average degree	Degree centrality	Eigenvector centrality	Common neighbor	Locality index
	0.067	0.061	0.26	0.13	0.095
Blockage rate	Average degree	Eigenvector centrality	Standard deviation of degree		Clustering coefficient
	0.063	0.41	0.081		0.076

show how we can distinguish users in terms of their social activity only by using the ensemble of their structure properties in the social network.

4 Evolutionary role mining

In the previous section, we observed that a variety of network characteristics show different degrees of correlation with the social roles of users regarding information propagation. In this section, we first introduce social role mining and use topological properties of users to infer their roles in a social event. The goal is to infer these roles in an information propagation process without knowing users activity and only using their position in the social network. The role of a user is not independent from its temporal behavior in the network, i.e., history of adding or removing links also affects its influence on the way information is propagated, as is depicted in Fig. 6. In particular, we examine the correlation between influence and the rate at which a user builds new connections over time. We define an user's degree growth rate as:

$$Growth\ rate(u_i) = \frac{degree^t(u_i) - degree^1(u_i)}{t} \tag{3}$$

where t is the age of user. As one can see from Fig. 6, influence score of a user i correlated to his temporal behavior. $Growthrate(u_i, t)$ is the difference between the initial degree of a user and its latest degree, normalized by its age i.e the amount of time elapsed since a user joined the network. Form Users with large growth rate means they attracted more connections per time step. For example, suppose user u_i joins the network at $t = 2$ with 10 edges and eventually has 40 connections at $t = 8$, then the $Growth\ rate(u_i) = (401 - 0)/6 = 5$ which means it has 5 new connections per time step. For another user u_j that joins the network at time $t = 6$ with 10 connections and has 40 connections at $t = 8$ we have $Growth\ rate(u_j) = (40 - 10)/2 = 15$. The neighborhood of u_j grew faster than neighborhood of u_i. From Fig. 6, we can see that influential users have large growth rates, meaning that they attract new connections faster. Social role mining is a dynamic process and here we introduce a new dynamic framework to categorize users into roles regarding their social relations and temporal behavior.

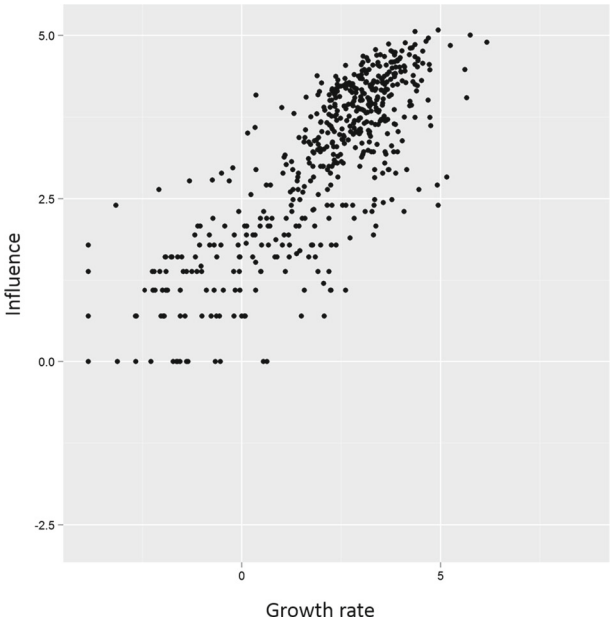


Fig. 6 The influence score and degree growth rate of users in Digg dataset. One pair in the plot shows influence of a user and the rate at which one user has built new connections over time

We first introduce the notations that we will use throughout the paper. We have a dynamic social network $G_t = (V_t, E_t, D_t)$ where $V_t = \cup_{i=1}^{t} V_i$ is the set of unlabeled users at time t, E_t represents the set of connections in the network and D_t is the set of structural properties of users. Suppose the set of labels $C_t = 1, \ldots, r$ represents the r clustering of users at time t and R_t is the social role of users in an information cascade.

The goal is to find the labels of users from their structural properties over time $\{D_1, \ldots, D_t\}$. We propose a dynamic ensemble clustering (Topchy et al. 2004; Fred and Jain 2002; Gionis et al. 2005; Strehl and Ghosh 2003) framework such that the partitioning of users represents the current role set in the network at time t and is also consistent with the historical information of users in previous time steps. In our method the clustering of D_t is derived from the aggregation of $C = \{C_1, \ldots, C_t\}$, a set of clusterings over time. We define a new similarity metric between users based on how similar they have been clustered over time. The partitioning of users based on this new metric gives the actual role of users at the current time.

The pseudo-code of our method is given in Algorithm 1. It takes as input: (1) G_t, the dynamic graph where edges are time stamped; (2) k, number of roles to extract; (3) $weightingfun(C)$, a weighting function to incorporate temporal behavior in partitioning; (4) $clusteralgo(simmatrix, k)$, an algorithm to partition users into k roles based on the calculated $simmatrix$ (a similarity matrix as detailed in Sect. 4.2).

Algorithm 1 Evolutionary Role Mining (ERM)

procedure ERM($G_T = (V, E_T), k, weightingfun(.), clusteralgo(.,.)$)
  for t in $1 : T$ **do**
    $D_t \leftarrow localProperties(G_t)$
    $C_t \leftarrow k - means(D_t, k)$
    $C \leftarrow C \cup C_t$
  end for
  $A \leftarrow weightingfun(C)$
  for t in $1 : T$ **do**
    $simmatrix \leftarrow simmatrix + pairwiseSimilarity(V, C_t) * \alpha_t$ ▷ $\alpha_t \in A$
  end for
  $R \leftarrow clusteralgo(simmatrix, k)$
  return R
end procedure

4.1 Local properties measurement

The first step in evolutionary role mining is to build clusters from structural properties of users D_t for all $t \in [1 : T]$. This clustering process is derived by applying the k-means clustering algorithm on D_t, where we minimize the euclidean distance between observations and centroids. We selected the structural properties of users that correlate to social roles of users, as explained in the previous section:

- the normalized node degree (K): quantifies the linkage of node i; it is the degree of node i divided by the sum of all nodes' degree in the network.
- the normalized average degree (r): shows the intensity of connectivity in the neighborhood of node i; it is calculated by averaging over all degree of immediate neighbors of node i.
- the standard deviation of degree (cv): coefficient variation of the degrees of the immediate neighbors of a node characterizes the coherence of the connectivity; it is measured by the standard deviation of the degrees in the neighborhood of node i.
- the clustering coefficient (cc): quantifies the connectivity between neighbors; it is measured as the proportion of existing connections between neighbors of node i to the number of all possible links between them (Watts and Strogatz 1998).
- the locality index (loc): characterizes the structure of neighbors' connectivity to the rest of the network; it is the ratio of links to the nodes outside of neighborhood to the number of links within the neighborhood to.
- the common neighbors (CN): measure the commitment of users to the neighborhood. This feature shows if neighborhood of a user has an overlap with its neighbors. It is the number of common neighbors between a user's direct connections.

$$CN_i = \sum_{u_j \in N_{u_i}} \frac{\left|N_{u_i} \cap N_{u_j}\right|}{\left|N_{u_i} \cup N_{u_j}\right|} \tag{4}$$

  where N_{u_i} is the set of neighbors of user i.
- the eigenvector centrality ($eig-cntr$): rank users regarding their importance in the network. This centrality measure acts similar to degree centrality however it gives higher score to the nodes which are themselves connected to high score nodes.

This feature vector has the advantage of measuring the connectivity of a node in its neighborhood structure and also it is fast to calculate. It has been shown that these properties can distinguish well nodes at different structural positions (Costa et al. 2009).

This step gives us a set of data clusters, $\{C_1, \ldots C_T\}$, which are then used to find the role of users at time T using evolutionary ensemble clustering.

4.2 Users similarity

We define the similarity matrix of users based on their co-clustering occurrence at previous time steps. The similarity matrix is an $n \times n$ matrix for n active nodes at the current time step. For two users u, v, if $C_i(u) = C_i(v)$ then $X_i(u, v) = 1$ and the total similarity of u, v is:

$$similarity(u, v) = \sum_{i=1}^{t} X_i(u, v) * \alpha_i \tag{5}$$

where X_i is an indicator function which tells if two users are in the same group in a clustering or not, and α_i is the weight of the clustering at time i. The value of α_i is determined by a weighting function.

4.3 Weighting functions

The network dynamics are embedded in our method by incorporating a weighting function which assigns more importance to some temporal data, and gives less weight to other data. We need a mechanism to identify those clusterings that are not consistent with the current clustering due to noise or concept drift. The common approach for these cases is to use either *temporal weighting*, or a *sliding window*. Another method for weighting the clustering is using the data distribution instead of the arrival time of data. We use two different scenarios to model the temporal behavior of data in clustering ensemble:

4.3.1 Temporal weighting

This function defines the probability that historic data is still valid for learning the roles at the current time. The basic idea of this weighting is that older data is less relevant to current data, so a lower weight is assigned to the older data. Different functions can be defined in this group but the general properties that all must hold are: (1) $0 \leqslant \alpha_i \leqslant 1$ for all $i \in [1, t]$; (2) $\alpha_i < \alpha_j, i < j$, (3) $\alpha_t = 1$.

Based on Cormode et al. (2009), we define a new exponential time decaying function, called temporal weighting (TW), to be used in our method:

$$\alpha_i = (1 - \theta)^{t-i}, \ for \ \ i = 1 \ \ to \ \ T. \tag{6}$$

4.3.2 Data distribution

The basic idea behind this type of weighting function is different from the previous one. Here, the validity of data is not assessed by its age but by its actual similarity to the current data. In this method, the older clustering that groups objects more similar to current data is more important than the recent clustering that does not. In other words, if $C_{t-k}(u) = C_t(u)$ and $C_{t-j}(u) \neq C_t(u)$ where $k > j$ then $\alpha_{t-k} > \alpha_{t-j}$ where C_t is the clustering at time t and α_t is the weight of clustering at time t. This method is used for weighting models to build ensemble classifiers (Wang et al. 2003).

We define data distribution weighting (DDW) function that assigns weight of data at each snapshot proportional to its similarity to the current data. We use the distance of two clusterings to define the weights. The distance of current clustering C_t and C_i is defined as the number of objects they have clustered differently (Gionis et al. 2005). The distance between two nodes u and v for two clusterings t and i is:

$$d_{u,v}(C_t, C_i) = \begin{cases} 1, & \text{if } C_t(u) = C_t(v) \text{ and } C_i(u) \neq C_i(v), \\ & \text{or } C_t(u) \neq C_t(v) \text{ and } C_i(u) = C_i(v) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

Then the distance of clusterings is measured as:

$$dist(C_t, C_i) = \sum_{u,v \in V_t} d_{u,v}(C_t, C_i) \tag{8}$$

As a consequence of the previous equation, the weight of clustering at time i is:

$$\alpha_i = 1 - Norm(dist(C_t, C_i)) \tag{9}$$

where $Norm(dist(C_t, C_i))$ is the value of distances normalized to the interval [0, 1].

4.4 Cluster ensembles

We obtain a set of clusterings of users for all time steps and combine these individual clusterings to obtain an ensemble clustering that categorizes users into social groups. We used three clustering algorithms on the derived similarity matrix derived from Eq. (5) to find the ensemble clustering.

We modified the *hypergraph partitioning algorithm* (*HGPA*) by Strehl and Ghosh (2003) to use the weighted similarity metrics. This method re-clusters the objects using the hyper-graph built upon the clusterings. In this method the hypergraph partitioning package HMETIS (Karypis et al. 1997) is used to partition the hyper graph.

Spectral clustering is tipically used for graph partitioning problems where a graph-based measure, such as the normalized cut, is to be minimized. This algorithm clusters objects based on the eigenvectors of their similarity matrix. For the nodes and their similarity, measured by Eq. (5), the graph Laplacian L is built: $L = S - W$ where S is the degree diagonal matrix of the similarity graph of nodes and W is the similarity

matrix of data. Then the first k eigenvectors of L are calculated. Finally the clustering is derived by applying k-means on a matrix, built from concatenation of the first k eigenvectors as columns (Von Luxburg 2007).

Another possibility for aggregating the clusterings over time is *agglomerative hierarchical (Agglo)* (Johnson 1967). This algorithm initially puts all objects in individual clusters then iteratively merges pairs of clusters either until deriving the defined number of clusters or until merging all the objects into one single cluster.

5 Experimental results

In this section we show the effectiveness of the proposed method on categorizing social roles of users in an information cascade. We applied the proposed methodology on two social networks from the Flickr and Digg datasets, described in Sect. 3.1. We evaluate the performance from two angles: (1) accuracy in categorizing users using F-score; (2) temporal consistency of the extracted role using normalized mutual information. We use the influence score and the blockage rate of users, as measured in Sects. 3.2 and 3.3, to define two sets of class labels as ground truth to compare our results against. These class labels are called "influence" and "blockage" and are respectively derived from equal width discretization of measured values of influence score and blockage rate where the number of intervals is 5.

5.1 Number of roles

We determine the number of roles by measuring F-score for different cluster sizes. We apply our method on the datasets for different cluster sizes from 5 to 25 and measure the F-score twice for each dataset: (1) F-score of results against influence categories as true labels; (2) F-score with blockage categories as the true labels. Hence we have 2 sets of results per dataset as shown in Fig. 7 for different number of roles. As we see, the best result (maximum F-score) is derived for cluster size 6 for Digg data and cluster size 8 for Flickr data. The F-score of p on r, denoted as $F(p, r)$, is the harmonic mean of precision and recall rates. For a predicted role p, we compute its F-score on each r in the actual roles of R and define the maximal obtained as p's F-score on R, i.e., $F(p, R) = \max_{r \in R} F(p, r)$. The final F-score of the predicted roles P on the actual roles

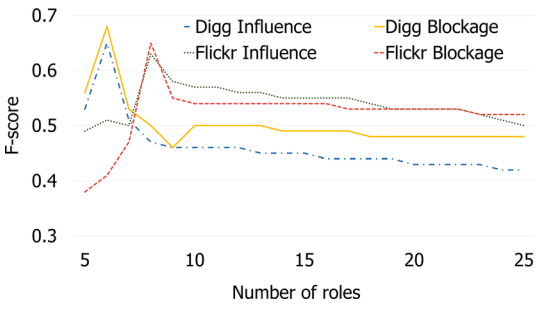


Fig. 7 Performance of derived roles by proposed method for different number of social roles in terms of F-score. The F-score is measured against two true class labels: influence and blockage rate (Color figure online)

R is then calculated as the weighted (by role size) average of each predicted role's F-score:

$$F(P, R) = \sum_{p \in P} \frac{|p|}{|V|} F(p, R) \tag{10}$$

For the predicted groups of p with reference to actual roles r (which are sets of nodes), the precision rate is defined as $\frac{|p \cap r|}{|p|}$ and the recall rate is defined as $\frac{|p \cap r|}{|r|}$. This effectively penalizes the predicted clustering that is not well aligned with the ground truth, and we use it as the quality measure of all methods on all datasets.

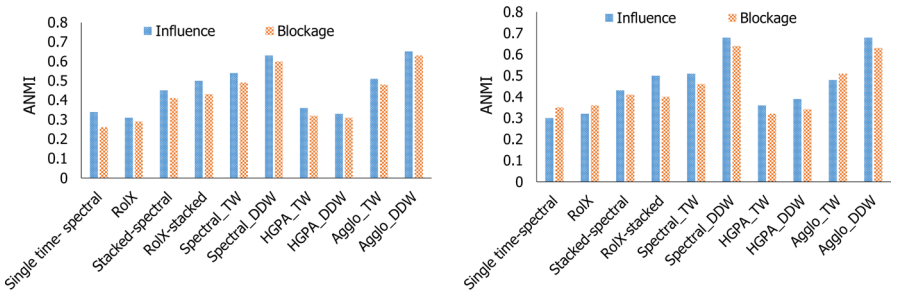
We varied the number of clusters from 5 to 25. If we assume that influence score and blockage rate are 100 percent correlated then there will be only 5 social groups, corresponding to the previously defined 5 equal-width groups. By contrast, if they are completely not correlated, there will be $5 \times 5 = 25$ groups. In practice, we found that the actual correlation between these two scores of users in Digg and Flickr social networks is respectively 0.13 and 0.24.

In this paper we used the F-score to evaluate the optimal number of clusters since we already have the desired labels of the users. In other applications, any method such as AIC (Akaike 1998) can be incorporated in our framework to determine the appropriate number of clusters.

5.2 Baseline methods

To show the effectiveness of our method, we compare the results of the proposed methodology against four baseline approaches. Since our method considers both the temporal behavior of users and their local structural properties, we use the following approaches to evaluate the performance of our method and study the effect of the clustering algorithm and the historical information:

- *Single time* This method studies the effect of historical data where only the local properties of users at the current time step are considered and the temporal behavior is discarded. In this method, we use k-means and spectral clustering to derive the social roles in the current snapshot of the network.
- *Stacked* In this method the temporal data is incorporated in the clustering by stacking all structural properties of users at each time step up to current time. The clusters are derived by applying a clustering algorithm on the stacked data, using k-means and spectral algorithms. This is the general approach in evolutionary clustering where all data is available. In addition, previous studies of dynamic role discovery employed this strategy (Choobdar et al. 2011; Rossi et al. 2012).
- *RolX* We also compare our method against the method proposed by Henderson et al. (2012). They use a set of structural features of nodes in networks and extract their role by applying a matrix factorization method to cluster nodes where each cluster represents a role. We extract the same feature vector as RolX, including local features, neighbor features and recursive features.
- *RolX-stacked* This method finds clusters of users by applying RolX method on stacked matrix of features where structural properties of users over time are aggregated into one matrix.



(a) Digg social network **(b)** Flickr social network

Fig. 8 Average normalized mutual information (ANMI) score for different approaches and proposed method (Color figure online)

5.3 Temporal consistency of social roles of users

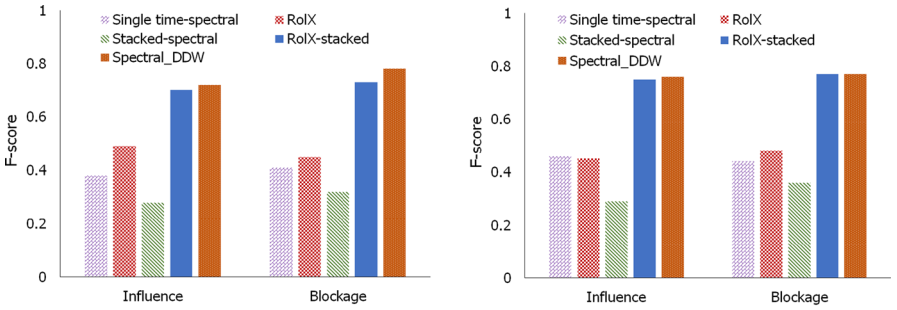
An important metric to evaluate the quality of the obtained social roles is the temporal smoothness of extracted roles. As we observed in Sect. 4, social roles of users in an information cascade are correlated to their history, hence the extracted roles should have consistency with previous time steps. We use the normalized mutual information (NMI) (Strehl and Ghosh 2003) measure to quantify the amount of mutual information shared between roles of users at previous time steps and current time.

Assume r groupings denoted as $R = \{\lambda^q, q \in \{1, \ldots, r\}\}$, then the normalized mutual information between two groupings λ^a and λ^b is estimated as:

$$\phi^{NMI}(\lambda^a, \lambda^b) = \frac{\sum_{h=1}^{k^a}\sum_{l=1}^{k^b} n_{h,l} log\left(\frac{n.n_{h,l}}{n_h^a n_l^b}\right)}{\sqrt{\sum_{h=1}^{k^a} n_h^{k^a} log\left(\frac{n_h}{n}\right)\sum_{l=1}^{k^b} n_l^{k^b} log\left(\frac{n_l}{n}\right)}} \tag{11}$$

where n_h^a is the number of objects in cluster C_h according to λ^a, and n_l^b the number of objects in cluster C_l according to λ^b, $n_{h,l}$ denote the number of objects that are in cluster h according to λ^a as well as in group l according to λ^b.

In Fig. 8 the performance of different weighting functions and algorithms on each dataset is compared against the baseline approaches. Each bar in this figure represents the average normalized mutual information (ANMI) between the set of r clusterings over time, R and the clustering at current time λ^T. In other words, we find the clustering of users for each timestamp t, then we compute the NMI relative to the clustering t and T, and we average the NMI values relative to all t. The performance of single time and stacked approaches for both clustering algorithms (k-means and spectral) are very similar. Thus, we only demonstrate the results of spectral clustering, in order to have the same base clustering algorithm for all methods, including ours, for a better and fairer comparison. The results of both datasets have a very similar pattern. We can see that our method outperforms the baseline approaches if the ensemble clustering algorithm is either spectral or agglomerative hierarchical clustering. HGPA has the worst performance for both weighting functions. Further investigation of data revealed that



(a) Digg social network **(b)** Flickr social network

Fig. 9 F-score performance of different approaches on the datasets (Color figure online)

the main reason for poor performance of HGPA was that it produces clusters with balanced sizes since it uses the HMETIS algorithm (Karypis et al. 1997). This algorithm partitions a graph with the constraint of producing even sized clusters. Whereas roles in a network are not equally distributed and some roles are at minority. The two other methods, spectral and agglomerative hierarchical clustering are at the same level of quality. This shows that the roles extracted by our method are more consistent over time and better show the dynamic of the network. The DDW weighting function produces better results in comparison to the temporal weighting function (TW). This function assigns more weight to the historic data that has a similar clustering structure to the current data. This suggests that social behavior of users is not monotone over time, hence the network structure at the current time is more similar to older times than just the previous snapshot of the network. In other words, if the topology of a network significantly changes over time, our method using DDW function can still find the structural roles of nodes with high accuracy (F-score) and consistency (NMI) including the concept drift in the structure of the network. By contrast, the two baseline methods suffer from this drawback: the stacked method uses the stacked dataset which is large and is likely to contain topological structure that is not valid for current snapshot; the single time method only considers one time step data which may not be enough for clustering. RolX also presents the same issue.

5.4 Performance in social role of users

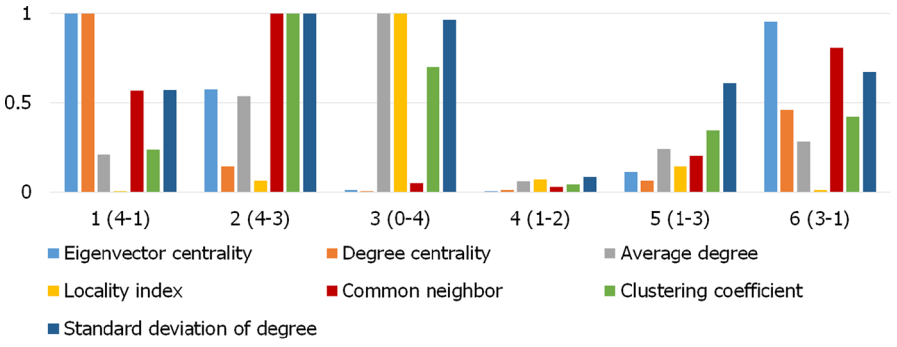
In this section we investigate how well the role of users correlate to their social influence and their function in information dissemination. We use the empirical influence rate and blockage rate of a user, as defined in Sects. 3.2 and 3.3, to label users to groups as the actual role of users in a cascade. We run our algorithm for Digg data for 6 roles and for Flickr data for 8 roles, as seen in Fig. 7. We compare the predicted social roles of users from our method with their actual roles (influence level and blockage rate) using F-score. Figure 9 shows the experiment results on Digg and Flickr data. In this figure we compare the performance of our methodology and the baseline approaches. For the sake of more clarity we only compare our methodology

using spectral clustering algorithm and the weighting function is data distribution since this configuration has better NMI performance over HGPA and relatively similar performance to agglomerative hierarchical clustering.

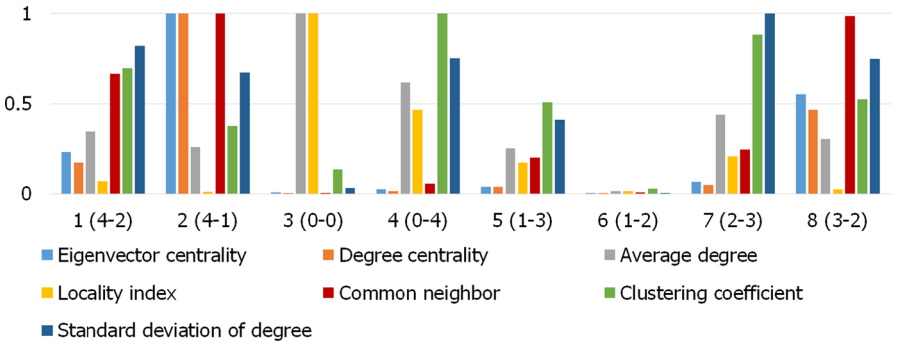
From the figures, we can see that the Spectral-DDW outperforms the baseline methods on F-score measures for both datasets of Digg and Flickr. Comparing to the single time approach, Spectral-DDW improves the performance by 0.3. This is in accordance with our observation that temporal behavior of users is important in categorizing users' social roles. This also suggests that the Spectral-DDW method is capable of incorporating history of users to infer their social roles. Moreover, it shows an improvement over the staked method. We can also observe that the performance of stacked method is lower than the single time approach. Due to the fact that this method uses all history of users, it is biased toward past. Hence, it does not reflect the current social roles of users accurately. Our proposed method outperformed RolX for both datasets. This can be due to the impact of temporal behavior of users in their social roles since the history of users is not considered in RolX. We also note that the Spectral-DDW method improves the F-score for the category of social roles on both the influence and blockage levels. This means that the Spectral-DDW proposed model is capable of categorizing users to their social roles in an information cascade. All these results hold for both datasets of Digg an Flickr, which shows the robustness and consistency of our method over different datasets. We can see that the results of RolX-stacked method and our proposed method are comparable. These experiments show that the quality of results by our method is not only due to incorporating temporal data but also is due to the method used for aggregating the history of users and their current status. In the stacked baseline approach, history is used and the base clustering is the same as in our method. However, the quality of results is lower than the quality of the roles discovered by the ensemble method we propose in this paper.

Figure 10 shows the average statistics for predicted roles in Digg and Flickr dataset. The x-axis shows the role number and each column represents the average value of local properties of users associated with predicted roles. For a better visualization, we normalized all the values to [0,1], so that we have the same range of values for all properties. We also report the category of users regarding their influence score and blockage rate for each predicted role. On the x-axis, the the numbers in the parentheses shows the category in form of (influence category–blockage category). These categories are the ones that have highest F-score with the predicted role. For both datasets we can see very similar patterns for grouping of users. Some of the roles are exactly the same in both dataset such as role 3 in Digg dataset and role 4 in Flickr dataset and the other have corresponding from one dataset to other such as role 2 in Digg and 1 in Flickr. We group the resulted roles based on their influence-role category where the numerical label of zero to four are translated to very low to very high in each category as follows:

– *Role A (very low influence, very high blockage)* The structural properties of this role are also very similar in both datasets. We can see that users with this role are located in a neighborhood with low cohesion as the variance of degree is high and reachability of users is low as their centrality measures are low and they are not committed to their neighbors, which also shows that level of trust in their



(a) Digg social network



(b) Flickr social network

Fig. 10 Average normalized local properties of users in predicted roles. The numbers on the *x* axis shows the predicted roles and corresponding social category in the form of (influence–blockage) (Color figure online)

neighborhood is low as they do not have much friends in common (Easley and Kleinberg 2012). Considering that these users have low influence and block many stories, this feature profile very well represents this role. This role corresponds to group 3 in Digg and group 4 in Flickr.

– *Role B (very high influence, low blockage)* Users with this role are very reachable and are located in a dense neighborhood with medium cohesion. As one can see from Fig. 10, commitment of users to their neighborhood is very high. Hence users are very influential and do propagate most of the stories they receive from their neighbors. This role includes group 1 from Digg and 2 from Flickr.
– *Role C (very low influence, very low blockage)* These users are not influential although they do propagate most of the received stories. This can be interpreted as this role belonging to users whose neighborhood is not well connected, but globally they are well connected as their locality index is very high. In addition, one can see that these users are connected to high degree users as their reachability features are low except for average degree. This role is only observed in Flickr dataset in group 3.

– *Role D (low-mid influence, high blockage)* Users in this group have low reachability but are connected to high degree users. Their neighborhood is nor as dense as users in role 2 and they are more committed to their neighbors than users in role 1. This role corresponds to group 5 in Digg and groups 5 and 7 in Flickr.
– *Role E (low influence, mid blockage)* Users in this role are not reachable nor committed to their neighbors. This structural position prevent them from being influential. In addition, their neighborhood is very coherent which means they are connected to users with similar degree which is low. This can explain why they are not very active users as they do not propagate many stories. This role corresponds to group 4 in Digg and group 6 in Flickr.
– *Role F (very high influence, mid-high blockage)* Users in this role are very influential but do not propagate most of the received stories. Structurally, they are very similar to role 2 except that they are not as reachable as users in role 2. This role corresponds to group 2 in Digg and group 1 in Flickr.
– *Role G (high influence, low-mid blockage)* What makes this group different from role 1 and role 6 is that they have high degree but are mostly connected to users of low degree. This role corresponds to group 6 in Digg and group 8 in Flickr.

5.5 Role generalization

In this section we examine the generalization power and the predictive ability of the discovered roles by applying our method to a classification task. We use the derived clusterings in the training set to classify the unlabeled users in the test set. We also use the cluster membership matrix of the users, which contains their distance to the centroids of the derived clusters. We exploit this matrix as a feature set with logistic regression to predict the role of users.

We also compare the predictive power of the discovered roles by our method (ERM) to the structural properties that ERM uses as input. We build a classifier (structReg) using logistic regression on the structural properties of users at the current time and temporal features including average, variance, min and max of structural features over time as input, instead of role membership matrix from ERM.

We divide the users in the network into 10 parts and, for each experiment, we select one section as test sets and run ERM on the other 9 sections as a training set to find the roles and build the cluster membership matrix (ninefold matrix). We build the classifier on ninefold matrix then predict the label of users in the test set (10th part of data which is not used by ERM). Here we use the labels of users derived from influence score

Fig. 11 The classification accuracy for Digg dataset over 10 different test set (Color figure online)

categorization in Sect. 3 as the class label. The results of this experiment on Digg dataset are shown in Fig. 11. The roles produced by ERM are capable to classify the users more accurately than structReg in the test set. (ERM = 78 %, structReg = 65 % average accuracy, p value = 0.012[3]). These results also show that ERM is capable to generalize more effectively than structReg, as we can see that the discovered roles on a part of a dataset can better classify users in the rest of the dataset.

6 Conclusions

In this paper we study online social networks and we try to infer social roles of users in information propagation based solely on their structural properties, and not on the information cascade itself. We divided users in five social roles under two categories of influence and blockage rate, which are two important characteristics of users regarding information propagating. We proposed a novel dynamic clustering model, which can integrate both the structural properties of individual users and their temporal behavior. The experimental results, using two real social network datasets, show that the proposed model greatly outperforms a number of baseline models and is able to effectively infer roles of users in an information cascade scenario. In our experiment, we have shown that the quality of the results obtained by our method is not only due to incorporating temporal data, but also due to the method used for aggregating the history of users and their current status.

In this study, we also explore the relation between users activity and their structural position. Among structural properties, we find that user commitment in the neighborhood has more impact on the influence score of users. In addition, neighborhood cohesion has a more additive effect on users' behavior in terms of blocking a cascade, and triadic closure is also useful. Our experiments show promising results in terms of correlation between user activities and their temporal structural properties and our model provides a step towards modeling an information cascade independently from the network of diffusion.

Acknowledgments Sarvenaz Choobdar is funded by an FCT Research Grant (SFRH/BD/72697/2010). Pedro Ribeiro is funded by an FCT Research Grant (SFRH/BPD/81695/2011). This work is partially funded by ERDF, COMPETE, and national funds through FCT within Project FCOMP-01-0124-FEDER-037281.

References

Adamic L, Adar E (2005) How to search a social network. Soc Netw 27(3):187–203

Agarwal N, Liu H, Tang L, Yu PS (2008) Identifying the influential bloggers in a community. In: Proceedings of ACM International Conference on Web Search and Data Mining

Akaike H (1998) Information theory and an extension of the maximum likelihood principle. Selected Papers of Hirotugu Akaike. Springer, New York

Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on twitter. In: Proceedings of ACM International Conference on Web Search and Data Mining

Bonacich P (2007) Some unique properties of eigenvector centrality. Soc Netw 29(4):555–564

[3] The p values are obtained from a one-tailed paired *t* test.

Cha M, Benevenuto F, Ahn Y-Y, Gummadi KP (2012) Delayed information cascades in flickr: measurement, analysis, and modeling. Comput Netw 56(3):1066–1076
Cha M, Haddadi H, Benevenuto F, Gummadi PK (2010) Measuring user influence in twitter: the million follower fallacy. In: Proceedings of AAAI International Conference on Weblogs and Social Media, vol. 10
Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the Flickr social network. In: Proceedings of ACM International Conference on World Wide Web
Choobdar S, Ribeiro P, Silva F (2012) Event detection in evolving networks. In: Proceedings of IEEE International Conference on Computational Aspects of Social Networks (CASoN), São Carlos, Brazil
Choobdar S, Silva F, Ribeiro P (2011) Network node label acquisition and tracking. In: Proceedings of Portuguese Conference on Artificial Intelligence, Progress in Artificial Intelligence
Cormode G, Shkapenyuk V, Srivastava D, Xu B (2009) Forward decay: a practical time decay model for streaming systems. In: Proceedings of IEEE International Conference on Data Engineering
Costa L, Rodrigues F, Hilgetag C, Kaiser M (2009) Beyond the average: detecting global singular nodes from local features in complex networks. Europhys Lett 87(1):18008
Danilevsky M, Wang C, Desai N, Han J (2013) Entity role discovery in hierarchical topical communities. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
Easley D, Kleinberg J (2010) Networks, crowds, and markets, vol 6(1). Cambridge University Press, New York
Easley D, Kleinberg J (2012) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, Cambridge
Fred A, Jain A (2002) Data clustering using evidence accumulation. In: Proceedings of International Conference on Pattern Recognition, vol. 4. Quebec City, Canada
Gallagher B, Eliassi-Rad T (2010) Leveraging label-independent features for classification in sparsely labeled networks: an empirical study. In: Proceedings of International Conference on Advances in Social Network Mining and Analysis. Berlin, Germany
Ghosh R, Lerman K (2010) Predicting influential users in online social networks. In: Proceedings of KDD Workshop on Social Network Analysis (SNA-KDD), July 2010
Ghosh R, Lerman K (2012) Rethinking centrality: the role of dynamical processes in social network analysis. arXiv preprint arXiv:1209.4616
Gionis A, Mannila H, Tsaparas P (2005) Clustering aggregation. In: Proceedings of IEEE International Conference on Data Engineering
Goyal A, Bonchi F, Lakshmanan LV (2010) Learning influence probabilities in social networks. In: Proceedings of ACM International Conference on Web Search and Data Mining
Granovetter M (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380
Granovetter M (1985) Economic action and social structure: the problem of embeddedness. Am J Sociol 91:481–510
Guo S, Wang M, Leskovec J (2011) The role of social networks in online shopping: information passing, price of trust, and consumer choice. In: Proceedings of the 12th ACM Conference on Electronic Commerce
Henderson K, Gallagher B, Eliassi-Rad T, Tong H, Basu S, Akoglu L, Koutra D, Faloutsos C, Li L, Matsubara Y, et al. (2012) Rolx: structural role extraction & mining in large graphs. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China
Iribarren JL, Moro E (2009) Impact of human activity patterns on the dynamics of information diffusion. Phys Rev Lett 103(3):038702
Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32(3):241–254
Karypis G, Aggarwal R, Kumar V, Shekhar S (1997) Multilevel hypergraph partitioning: application in vlsi domain. In: Proceedings of the 34th Annual Design Automation Conference, ACM
Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. Science 311(5757):88–90
Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of ACM International Conference on World Wide Web
Lancichinetti A, Fortunato S (2012) Consensus clustering in complex networks. Sci Rep 2:336
Lee C, Kwak H, Park H, Moon S (2010) Finding influentials based on the temporal order of information adoption in twitter. In Proceedings of ACM International Conference on World Wide Web

Lerman K, Ghosh R, Surachawala T (2012) Social contagion: an empirical study of information spread on digg and twitter follower graphs. arXiv preprint arXiv:1202.3162
Myers SA, Zhu C, Leskovec J (2012) Information diffusion and external influence in networks. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
Rissanen J (1978) Modeling by shortest data description. Automatica 14(5):465–471
Romero DM, Galuba W, Asur S, Huberman BA (2011) Influence and passivity in social media. In: Proceedings of the ECML/PKDD
Rossi R, Gallagher B, Neville J, Henderson K (2012) Role-dynamics: fast mining of large dynamic networks. In: Proceedings of ACM International Conference on World Wide Web. Lyon, France
Saito K, Nakano R, Kimura M (2008) Prediction of information diffusion probabilities for independent cascade model. Knowledge-based intelligent information and engineering systems. Springer, New York
Strehl A, Ghosh J (2003) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3:583–617
Tang J, Sun J, Wang C, Yang Z (2009) Social influence analysis in large-scale networks. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
Taskar B, Abbeel P, Koller D (2002) Discriminative probabilistic models for relational data. In: Proceedings of Conference on Uncertainty in Artificial Intelligence. Alberta, Canada
Topchy A, Law M, Jain A, Fred A (2004) Analysis of consensus partition in cluster ensemble. In: Proceedings of IEEE International Conference on Data Mining. Brighton, UK
Ver Steeg G, Ghosh R, Lerman K (2011) What stops social epidemics? In: Proceedings of AAAI International Conference on Weblogs and Social Media
Von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416
Wang H, Fan W, Yu PS, Han J (2003) Mining concept-drifting data streams using ensemble classifiers. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
Wang T, Srivatsa M, Agrawal D, Liu L (2012) Microscopic social influence. In: Proceedings of SIAM International Conference on Data Mining
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442
Zhao Y, Wang G, Yu PS, Liu S, Zhang S (2013) Inferring social roles and statuses in social networks. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
Zhou Y, Liu L (2013) Social influence based clustering of heterogeneous information networks. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining