CrossMark

ORIGINAL ARTICLE

Homophilic network decomposition: a community-centric analysis of online social services

Giulio Rossetti[1,2] · Luca Pappalardo[2] · Riivo Kikas[3] · Dino Pedreschi[2] ·
Fosca Giannotti[1] · Marlon Dumas[3]

Abstract In this paper we formulate the *homophilic network decomposition* problem: Is it possible to identify a network partition whose structure is able to characterize the degree of homophily of its nodes? The aim of our work is to understand the relations between the homophily of individuals and the topological features expressed by specific network substructures. We apply several community detection algorithms on three large-scale online social networks—Skype, LastFM and Google+—and advocate the need of identifying the right algorithm for each specific network in order to extract a homophilic network decomposition. Our results show clear relations between the topological features of communities and the degree of homophily of their nodes in three online social scenarios: product engagement in the Skype network, number of listened songs on LastFM and homogeneous level of education among users of Google+.

✉ Giulio Rossetti
  giulio.rossetti@isti.cnr.it

  Luca Pappalardo
  lpappalardo@di.unipi.it

  Riivo Kikas
  kikas@ut.ee

  Dino Pedreschi
  dino.pedreschi@di.unipi.it

  Fosca Giannotti
  fosca.giannotti@isti.cnr.it

  Marlon Dumas
  marlon.dumas@ut.ee

[1] KDDLab, ISTI -CNR, Via G. Moruzzi, 1, 56124 Pisa, Italy

[2] KDDLab, University of Pisa, Largo B. Pontecorvo, 3, 56127 Pisa, Italy

[3] Unversity of Tartu, Tartu, Estonia

1 Introduction

As the social media space grows more and more people interact and share experiences through a plethora of different online services, producing every day a huge amount of personal data. Companies providing social media platforms are interested in exploiting these Big Data to understand "user engagement," i.e., the way individuals use products provided via the platform. In particular predictive analytics allows these companies to exploit historical user engagement data, in conjunction with social network data, in order to predict future product usage (engagement) of individuals in the network. Traditional approaches of predictive analytics focus on *individuals*: they try to describe and predict the level of engagement of a single individual. Focusing on individuals, however, introduces many challenging issues. First the amount of individuals to process is enormous and hence hardly manageable. Think about online giants like Skype or Facebook: in these contexts providing an up-to-date description and prediction of user engagement for billion of users is not practically feasible. Second addressing each single individual is in many cases redundant, since neighbors in networks tend to behave in a similar way and to share specific features (age, location, language, interests), i.e., they show a certain degree of social homophily (McPherson et al. 2001; Himelboim et al. 2013). Indeed the analysis of user engagement can be seen as an instantiation of a more general problem: *homophilic network decomposition*, which consists in finding a partition of the network which guarantees a high degree of homophily in the subgroups of the network.

Restricting the analysis to single users inevitably causes the underestimation of the importance of social homophily, whereas online social services are usually designed to

foster social interactions between individuals. It is hence fundamental to widen the analysis spectrum in order to incorporate the social surrounding of users in order to capture the homophily which characterizes real social networks. We propose to move the focus from individuals to groups, i.e., to analyze and describe the level of homophily of *social communities*. If user-centric approaches fail because they do not take into account the individuals' social surroundings, on the other hand, it goes without saying that analyzing the homophily on the overall network does not make sense. The group-centric approach focuses on social communities as a trade-off between the micro- and the macro-level of network granularity (Fig. 1). Moving the interest from individuals to communities brings many advantages. First we reduce by several orders of magnitude the space of analysis, shrinking the number of objects to process and speeding up the analytical tasks. Second targeting communities allows for capturing the homophily inherent to the social network: we can "compress" into one object all the densely connected components of a social group. Finally groups are complex objects from which we can extract a wide set of features for the analysis.

In this paper we investigate the potential of a group-centric approach in describing user homophily. Using different community detection algorithms we compute social communities from three large-scale online social networks (Skype, LastFM and Google+) and extract salient features from each community. We then build a repertoire of classifiers to predict the level of homophily in the communities both in terms of product engagement and similarity of attributes. We find two main results. First group-centric approaches outperform user-centric ones when we use algorithms producing overlapping micro-communities. In contrast, adopting partitioning algorithms which maximize modularity and produce macro-communities, the performances are worse than the ones of classical user-centric strategies. Second the group-centric approach is useful when dealing with networks where social interactions are a crucial part of the online service, such as the Skype social network, while it fails when the social network is just a marginal part of the service, such as for LastFM. Our work shows how the choice of a proper community detection algorithm—for the specific network analyzed—is crucial to partition the network into

Fig. 1 Interpolation between the local and the global level through network partitions of different sizes

homophilic groups of users. Moreover, varying the online social services analyzed (and related semantics) we observe that the obtained communities are proxies for the homophily in the network.

The rest of the paper is structured as follows. Section 2 defines the problem of homophilic network decomposition, which is the basis for addressing predictive tasks at the level of groups of individuals. Section 3 introduces the datasets and experimental setup used to test different methods to address the problem of homophilic network decomposition. Section 4 presents the experimental results, while Sect. 5 discusses the implications of these results. Finally Sect. 6 discusses related work and Sect. 7 provides a summary of the contribution suggesting directions for future work.

2 Problem definition

Online social services enable people to share interests, interact and generate content. The users of these services naturally tend to cluster around similar attributes (i.e., age, location and tastes), a property called *social homophily* (McPherson et al. 2001). To identify homophilic behaviors we need to identify the right observation granularity: Which is the subgraph size that maximizes the similarity of users w.r.t. to a given attribute? Are specific online social networks more homophilic than others? These questions are instantiations of the more general problem of *homophilic network decomposition*:

Definition 1 (*Homophilic network decomposition*) Given a social graph $G = (V, E)$ and a set L of node labels, an *homophilic network decomposition* is a collection of subgraphs of G, i.e., $H = \{G_1, \ldots, G_n\}$ where $G_1 = (V_1, E_1)$, $\ldots$, $G_n = (V_n, E_n)$, such that $\forall i \in [1..n], V_i \subset V \wedge E_i \subset E$ and in each subgraph G_i there is a dominant label, i.e., $\forall G_i \exists l \in L \mid \frac{|\{v \in V_i | L(v) = l\}|}{|V_i|} > \tau$. In this context, τ is the *dominance threshold*, meaning that the proportion of nodes in G_i that have the dominant label is at least τ.

One key question in *homophilic network decomposition* is how to break down the network in a way that is topologically meaningful and preserves the desired homophily property inside each group.

In this work we address the problem of *homophilic network decomposition* in three different online social networks: the full Skype contact graph, a nationwide Google+ snapshot and a sample of UK users of the LastFM social network. All these networks have peculiar structures, node attributes and semantics: we select for each network a target feature (not directly related with network topology) and identify the best partition across a set of candidates as a

classification problem. Our aim is to measure the ability of the topology of a community in estimating the homophily of nodes in the community: Do users in dense LastFM communities listen to more music? Do users in big Google+ communities have the same education level? Are Skype users in nation- homophilic communities frequent video callers?

Our experiments address these questions, and the obtained results are used to discuss the differences among the analyzed networks and the role their semantics play in the quality of the classification results.

3 Experimental settings

In this section we define our experimental settings: in Sect. 3.1 we describe the three online social network datasets we analyze, and in Sect. 3.2 we present the community discovery algorithms we use to partition the networks. Finally in Sect. 3.3 we introduce the topological features used to train the classifiers that discriminate high from low homophilic communities.

3.1 Datasets description

We analyze three large-scale datasets of popular online platforms: Skype, LastFM and Google+.

3.1.1 The Skype dataset

The first dataset is provided by Skype and includes anonymized data of Skype users as of October 2011. Each user (identified by hashed identifier) is associated with an account creation date, a country, and city of account creation. The dataset also includes undirected connections between users: a connection exists between two users if and only if they belong to each other's contact list. Connections are established as follows: If a user u wants to add another user v to her contact list, u sends v a contact request. The connection is established at the moment v approves the request (or not established if the contact request is not approved). In the dataset, each connection is labeled with a timestamp corresponding to the contact request approval. The dataset also includes data about usage of two Skype products: video calling and chatting. Product usage is aggregated monthly. Specifically, for each product, for each user and for each month, we are given the number of days in the month when the user used the product in question. The product usage data do not provide information about individual interactions between users, such as participants in an interaction, content, length or time of the interaction. The frequency of product usage is not recorded at a finer granularity than monthly. In this

paper, we focus on analyzing the most recent available snapshot of the network. Accordingly we focus on the subset of the dataset containing only users who used one of the two products, during at least two of the last three months covered in the dataset. Our analyses are then executed on a filtered dataset composed by several tens of millions of users and connections.

3.1.2 The LastFM dataset

LastFM is a popular online social network platform where people can share their music tastes and discover new music based on what they like. Once a user subscribes to an account, she can either start listening LastFM personalized Radio or send data about her own offline listenings. For each song a user can express her preferences and add tags (e.g., genre of the song). Lastly a user can add friends (undirected connections, the friendship request must be confirmed) and search for other users with similar musical tastes. A user can see, in her homepage, her friends' activities. Using LastFM APIs[1] we downloaded a sample of the UK user graph, starting from a set of nodes and implementing a breadth-first approach. We decided to explore the graph up to the fifth degree of separation from our seeds. For each user, we retrieved: (a) her connections and (b) for each week in the time window from January 2010 to December 2011, the number of single listenings of a given artist (e.g., in the first week of April 2010, user 1324 has listened 66 songs from the artist Bon Jovi). The number of listenings gives an estimate of the engagement of the user with respect to the LastFM service. Each song has a tag representing the music genre of the song (rock, metal, jazz, punk, etc.). After the crawl and cleaning stages, we build a social network where every node is a user and each edge is generated by looking at the user's friends in the social media platform. The total amount of nodes is 75, 969, with 389, 639 edges connecting them.

3.1.3 The Google+ dataset

Google+ is an interest-based social network that is owned and operated by Google. Each user in Google+ has a public visible account and can create links with other users inserting them in proper social circles. In this paper we use a social network built on the Google+ service upon US users, crawled by authors of Gong et al. (2012). Each user has also attached semantic information about education level, i.e., node labels identifying the schools attended by the users. The network contains 33,381 nodes and 110,142 edges.

---

[1] http://www.last.fm/api.

3.2 Community detection algorithms

Among the many different community detection algorithms proposed so far we identify two archetypal classes: the algorithms that maximize community *density* and the ones that maximize *modularity*. The former class ensures a high density of links inside communities, while the latter class imposes that the density of links inside a community is higher than the density of links which connect a community to external nodes. The degree of *overlap* is another property that discriminates between community discovery (henceforth, CD) algorithms. Classical approaches produce a partition of the network, i.e., an individual can be involved in at most one community. Overlapping approaches consider instead the multidimensional nature of social networks allowing the individuals to belong to many different communities.

We use four different algorithms to extract social communities from the Skype network: LOUVAIN, HDEMON, EGO-NETWORK and BFS. Such algorithms cover several declinations of both overlap and density/modularity optimization.

LOUVAIN Blondel et al. (2008) is a fast and scalable algorithm based on a greedy modularity approach. It performs a modularity optimization in two steps. First the method looks for "small" communities by optimizing modularity locally. Second it aggregates nodes belonging to the same community and builds a new network whose nodes are communities. These steps are repeated iteratively until a maximum of modularity is obtained, producing a hierarchy of communities. LOUVAIN produces a complete non-overlapping partitioning of the graph. It has been shown that modularity-based approaches suffer a resolution limit, and therefore, LOUVAIN is unable to detect medium size communities (Fortunato and Barthélemy 2007). This produces communities with high average density, due to the identification of a predominant set of very small communities (usually composed by 2–3 nodes) and a few huge communities. The LOUVAIN algorithm, which is parameter-free, produces a hierarchy of seven levels when applied on the Skype dataset.

HDEMON Coscia et al. (2014) is based on a recursive aggregation of denser areas extracted from ego-networks. Its definition allows to compute communities with high internal density and tunable overlap. In its first hierarchical level HDEMON operates extracting ego-networks and partitioning them into denser areas using label propagation. The communities computed at a given hierarchical level are subsequently used as meta-nodes to build a new network in the next hierarchical level, where the edges between the meta-nodes are weighted using the Jaccard of meta-nodes' contents. This procedure stops when disconnected meta-nodes, identifying the components of the original network, are obtained. The algorithm has two parameters: (1) the minimum community size μ and (2) the minimum Jaccard ψ among meta-nodes to create an edge that connects them. We apply HDEMON on the Skype dataset fixing $\mu = 3$ (the minimum community is a triangle) and using two different values of the ψ parameter: $\psi = 0.25$ which produces the HDEMON25 community set, and $\psi = 0.5$ which produces the HDEMON50 community set. For each community set we consider only the first 5 levels of the produced community hierarchy.[2] For LastFM and Google+ we only use the first hierarchical level produced by HDEMON, because of the reduced sizes of the datasets.

EGO-NETWORK is a naive algorithm that models the communities as the set of induced subgraphs obtained considering each node with its neighbors. This approach provides the highest overlap among the four considered approaches: each node u belongs exactly to $|\Gamma(u)| + 1$ communities, where $\Gamma(u)$ identifies its neighbors set. We apply a node sampling strategy and consider only a ratio ϵ of the ego-networks for the analysis. We set the parameter $\epsilon = 0.2$ and randomly extracted a number of users equals to the 20 % of the population. We choose $\epsilon = 0.20$ because it produces a community overlap similar to the one produced by HDemon. For each random user we extracted the corresponding ego-network, filtering only unique ones (two users can have equal ego-networks if they share all their contacts).

The BFS algorithm extracts randomly connected components from the graph. It randomly samples a ratio ϵ of the nodes of the network and, for each one of them, a number *csize* is extracted from a power law distribution of community sizes. Similarly to EGO-NETWORK, we choose $\epsilon = 0.20$. As parameters for the power law distribution of community size we choose the exponent $\beta = 1.8$ and the cutoff $\tau = 10,000$, which are the values we observe for HDemon25 on the Skype dataset.[3] Starting from a root node, the algorithm explores other nodes performing a breadth-first search and stopping when *csize* nodes are discovered.

Both HDEMON and LOUVAIN generate different community sets at different granularity, according to the parameters. For the Skype network, due to its size, we choose to analyze the two levels of the HDEMON hierarchy having the highest average community density and the community set

---

[2] We report the results of HDemon for $\psi = 0.25$ and $\psi = 0.50$ only. For $\psi < 0.25$ (i.e., low Jaccard in merge) there is an increase in network density which produces a small number of huge communities, similarly to LOUVAIN. For $\psi > 0.50$ (i.e., high Jaccard in merge) we obtain an incomplete node coverage, i.e., most of the nodes in the network are not assigned to a community.

[3] We observe similar values of β and τ on HDemon25 and HDemon50 on the Skype, LastFM and Google+ datasets.

at level 0 and 6 for LOUVAIN, which corresponds, respectively, to the first greedy iteration and the iteration having the maximum modularity. Conversely for the analysis of LastFM and Google+ we consider only the first hierarchical level produced by HDEMON (we will refer to as DEMON) and the last level of LOUVAIN which guarantees the maximum modularity. Also on LastFM and Google+ we do not apply BFS due to their reduced size.

3.3 Community feature extraction

From the community sets produced by the four algorithms we extract a set of *structural* features (see Table 1), which convey information about the topology of a social community $C = (V_C, E_C)$, where V_C and E_C are the set of nodes and edges in the community, respectively. The number of nodes N and edges M provides information about the community size. The community density $D = \frac{2M}{N(N-1)}$, i.e., the ratio between the actual links and all the possible links, indicates the level of interaction within the social group. The clustering coefficient (Watts and Strogatz 1998) indicates how strong is the presence of triangles within the community, measuring a "all-my-friends-know-each-other" property. The degree assortativity A_{deg} indicates the preference for the nodes to attach to others that have the same degree (Newman 2003). Other structural features regard the level of hubbiness of a community, such as the average/maximum degree computed considering both the network links or the community links only. The diameter $d = \max_{v \in V} \epsilon(v)$ and the radius $r = \min_{v \in V} \epsilon(v)$ are, respectively, the maximum and the minimum eccentricity ϵ of any node, where the eccentricity $\epsilon(v)$ is the greatest geodesic distance between a node v and any other node in the community. They represent the linear size of a community. Finally other structural features are considered, such as the number of community neighborhoods (nodes in the global network connected to nodes in the community), the number of edges leaving the community, the number of triangles and the number of connected triples.

Moreover, for the Skype dataset we introduce two additional feature sets: *community formation* features and *geographical* features (see Table 2). The *community formation* features convey information regarding the temporal appearance of nodes within the community, such as the time of subscription to Skype of the first user to subscribe; the average and the standard deviation of the inter-arrival times of users; the inter-arrival time between the first node to subscribe and the last node who adopted Skype. *Geographical* features provide information about the geographical diversity of a community or, in other words, its cosmopolitan nature. The number of different countries represented gives a first estimation of the international nature of the community. The country entropy estimates the national diversity through the Shannon entropy: $E = \sum_{c \in C} p(c) \log p(c)$, where C is the set of the countries represented in the community and $p(c)$ is the probability of the country c to be represented in the community. We also compute the city entropy and the number of different cities represented by the community. Moreover, for the users for which we know the city name (those associated with cities with more than 5000 Skype users), we compute their geographical distance using the coordinates of the centers of the cities. Once computed all the available distances, we consider the average and the maximum geographical distances of each community.

Finally for each network we define the target features we want to predict using the topological (and formation/geographical) features. For Skype the target features indicate the mean level of Skype activity performed by the community members. For such dataset we extract two target features: (1) *chat*, the mean number of days they used the instant messaging (chat) and (2) *video*, the mean number of days they used the video conference. Conversely the LastFM target feature indicates the mean level of user listening activity (i.e., the average of the number of listenings among the users of each community) while in Google+ it identifies the homogeneity of the users w.r.t. the education level (computed through node label entropy).

Table 1 Description of the structural features extracted from the communities

Structural features			
N	Number of nodes	M	Number of edges
D	Density	CC	Global clustering
CC_{avg}	Average clustering	A_{deg}	Degree assortativity
$\deg_{max}^{C}$	Max degree (community links)	$\deg_{avg}^{C}$	Avg degree (community links)
$\deg_{max}^{all}$	Max degree (all links)	$\deg_{avg}^{all}$	Avg degree (all links)
T	Closed triads	T_{open}	Open triads
O_v	Neighborhood nodes	O_e	Outgoing edges
E_{dist}	Num. edges with distance	d	Approx. diameter
r	Approx. radius	g	Conductance

Table 2 Description of the community formation features and geographical features extracted from the communities (only for the Skype dataset)

Community formation features	
T_f	First user arrival time
IT_{avg}	Avg user inter-arrival time
IT_{std}	Std of user inter-arrival time
$IT_{l,f}$	Last–first inter-arrival time
Geographical features	
N_s	Number of countries
E_s	Country entropy
S_{max}	Percentage of most represented country
N_t	Number of cities
E_t	City entropy
$dist_{avg}$	Avg geographical distance
$dist_{max}$	Max geographical distance

4 Analytical results

In this section we construct the classification models to estimate the degree of homophily from the community features. In Sect. 4.1 we start with the Skype contact graph describing a specific instantiation of the analyzed problem, namely *Social Engagement*. In such scenario we are interested in using topological, geographical and temporal network features to estimate the average engagement each community has on two Skype products, video and chat. In Sect. 4.2 we analyze the LastFM graph and shift our attention on a different formulation of our original problem: *Service Engagement*. Here, we want to estimate the average community level of music listening, i.e., how much users in the same community use in average the LastFM scrobbler (estimated by the gross number of her listenings). Finally in Sect. 4.3 we address the problem of estimating the degree of homophily w.r.t. the education level of Google+ users within communities.

4.1 Skype: user engagement

We use the topological, geographical and temporal features described above to *classify* the level of engagement of social communities with respect to the chat and video activity features. To this purpose, we build a supervised classifier that assigns communities to two possible categories: high level of engagement or low level of engagement. We address two different scenarios: (1) a balanced class scenario where the two classes have the same percentage of population and (2) an unbalanced class scenario, where we consider an uneven population distribution.

4.1.1 Balanced scenario

We consider two classes of user engagement for each of the two activity features (chat and video): low engagement and high engagement. To transform the two continuous activity features into discrete variables we partition the range of values through the median of their distribution. This produces, for each variable to predict, two equally populated classes: (1) low engagement, ranging in the interval [0, *median*] and (2) high engagement, ranging in the interval [*median*, 31].[4] To perform classification we use stochastic gradient descent (SGD) and area under the ROC curve (AUC) to evaluate their performance. The ROC curve illustrates the performance of a binary classifier and is created by plotting the true positive rate (tpr, also called sensitivity) versus the false positive rate (fpr, also called fallout or 1-specificity), at various threshold settings. The overall accuracy is instead the proportion of true results (both true positives and true negatives) in the population. Moreover, in a preliminary testing phase the classification step was repeated also using a random forest model built upon C4.5: due to the similar performance observed, the more intuitively interpretation of the obtained results and the lower execution time we decided to show only the results obtained by SGD.

We learn the SGD classifier with logistic error function (Tsuruoka et al. 2009; Zhang 2004) exploiting its implementation provided by the `sklearn` Python library.[5] We execute 5 iterations, performing data shuffling before each one of them, imposing the elastic-net penalty $\alpha = 0.0001$ and $l1$-ratio = 0.05. The adoption of elastic-net penalty results in some feature weights set to zero, thus eliminating less important features.

We apply a fivefold cross-validation for learning and testing. Table 3 shows the AUC produced by the SGD method on the features extracted from the community sets produced by the four algorithms (for HDEMON and LOUVAIN only the two best performing community sets are reported). HDEMON produces the best performance, both in terms of AUC and overall accuracy, for all the three activity features. LOUVAIN, conversely, reaches a poor performance, and it is outperformed by the more trivial BFS and EGO-NETWORK algorithms. This result suggests that the adoption of modularity optimization approaches, like LOUVAIN, is not effective when categorizing group-based user engagement due to their resolution limit which causes the creation of huge communities (Fortunato and Barthélemy 2007). As the level of the LOUVAIN hierarchy increases, and hence, the modularity increases, both the AUC and overall accuracy

---

[4] The maximum is 31 because it refers to the mean number of days per month in which that activity was performed.

[5] http://scikit-learn.org/stable/index.html

Table 3 Skype: AUC and accuracy (within brackets) produced by the SGD method in the balanced scenario, for video and chat features

| Algorithm	Lv.	Scores
Video: AUC and accuracy		
HDEMON25	1	**.74** (.67)
HDEMON50	0	.71 (.68)
LOUVAIN	0	.65 (.60)
LOUVAIN	6	.63 (.59)
EGO-NETS	–	.70 (.64)
BFS	–	.67 (.62)
Chat: AUC and accuracy		
HDEMON25	2	**.84** (.77)
HDEMON50	1	.81 (.73)
LOUVAIN	0	.69 (.64)
LOUVAIN	6	.65 (.60)
EGO-NETS	–	.75 (.75)
BFS	–	.81 (.72)

In bold the best model

decrease. In the experiments, indeed, the first LOUVAIN hierarchical level outperforms the last level, even though the latter has the highest modularity. Figure 2 shows the features which obtain a weight value by the SGD method higher than 0.2 or lower than −0.2 (i.e., the most

discriminative features for the classification process). HDEMON distributes the weights in a less skewed way, while the other algorithms tend to give high importance to a limited subset of the extracted features. Moreover only a few LOUVAIN features have a weight higher than 0.2 or lower than −0.2 (see Fig. 3d), confirming that a modularity approach produces communities with weak predictive power with respect to user engagement. Moreover, an interesting phenomenon emerges: independently from the chosen community discovery approach, the most relevant class of features for the classification process seems to be to the *topological* one (i.e., the sum of the absolute values of the SGD weights for the features belonging to such class is always greater than the same sum for *community formation* and *geographical* features combined). In particular degree, density, community size and clustering-related measures often appear among the most weighted features. Figure 4 shows the relationships between the average community size, the average community density and the AUC value produced by the SGD method on the community sets which reach the best performances in the balanced scenario. The best performance is obtained for the HDEMON community sets, which constitute a compromise between the micro- and the macro-level of network granularity.

Fig. 2 Skype: weights of the features (>|0.2|) produced by the SGD method for each community set for the chat feature in the balanced scenario. **a** HDEMON chat. **b** EGO-NETWORK chat. **c** BFS chat. **d** LOUVAIN chat
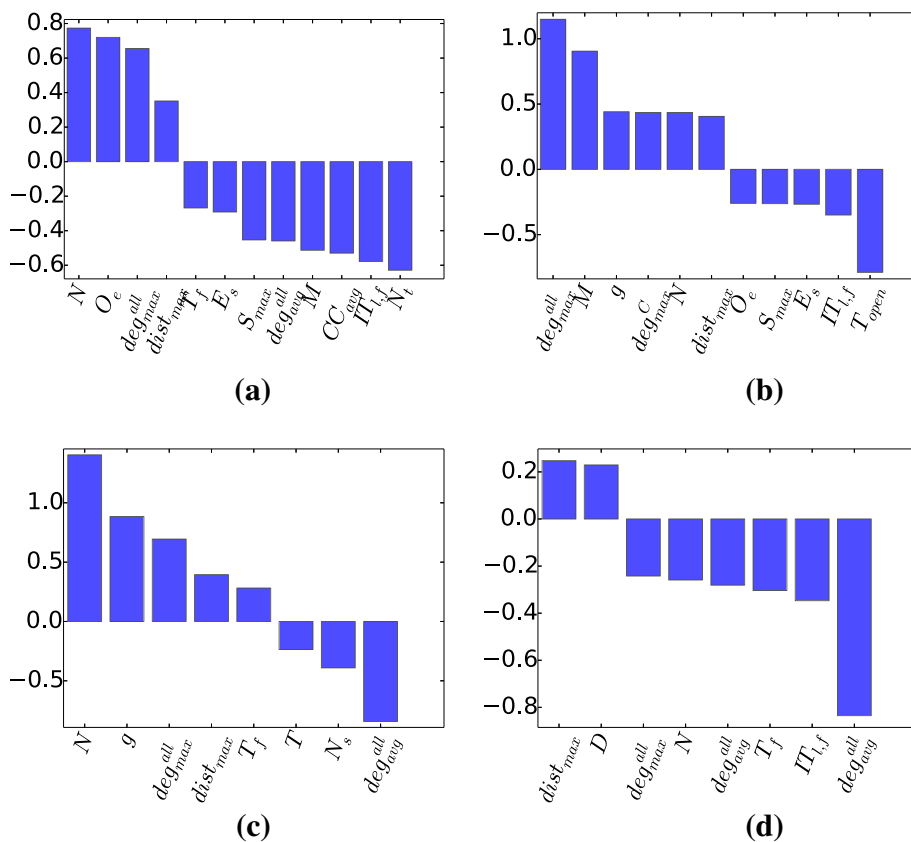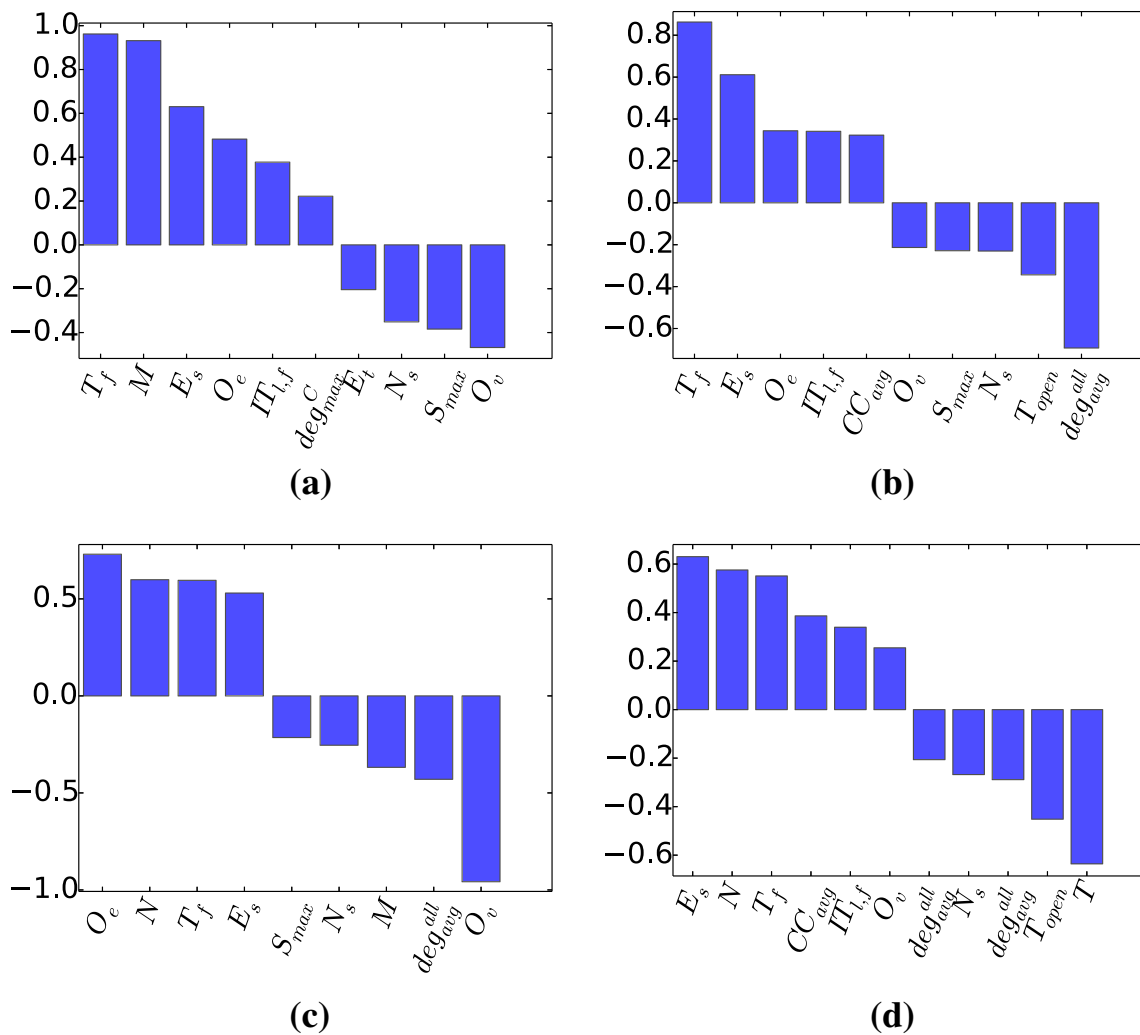
Fig. 3 Skype: weights of the features produced by the SGD method for each community set for the video feature in the unbalanced scenario. **a** HDEMON video. **b** EGO-NETWORK video. **c** BFS video. **d** LOUVAIN video

When the average size of the communities is too low, as for the ego-network level, we lose information about the surroundings of nodes and do not capture the inner homophily hidden in the social context. On the other hand, when communities become too large, as in the case of communities produced by LOUVAIN we mix together different social contexts losing definition. Communities expressing a good trade-off between size and density, as in the case of the HDEMON algorithm, effectively reach the best performance in the problem of estimating user engagement.
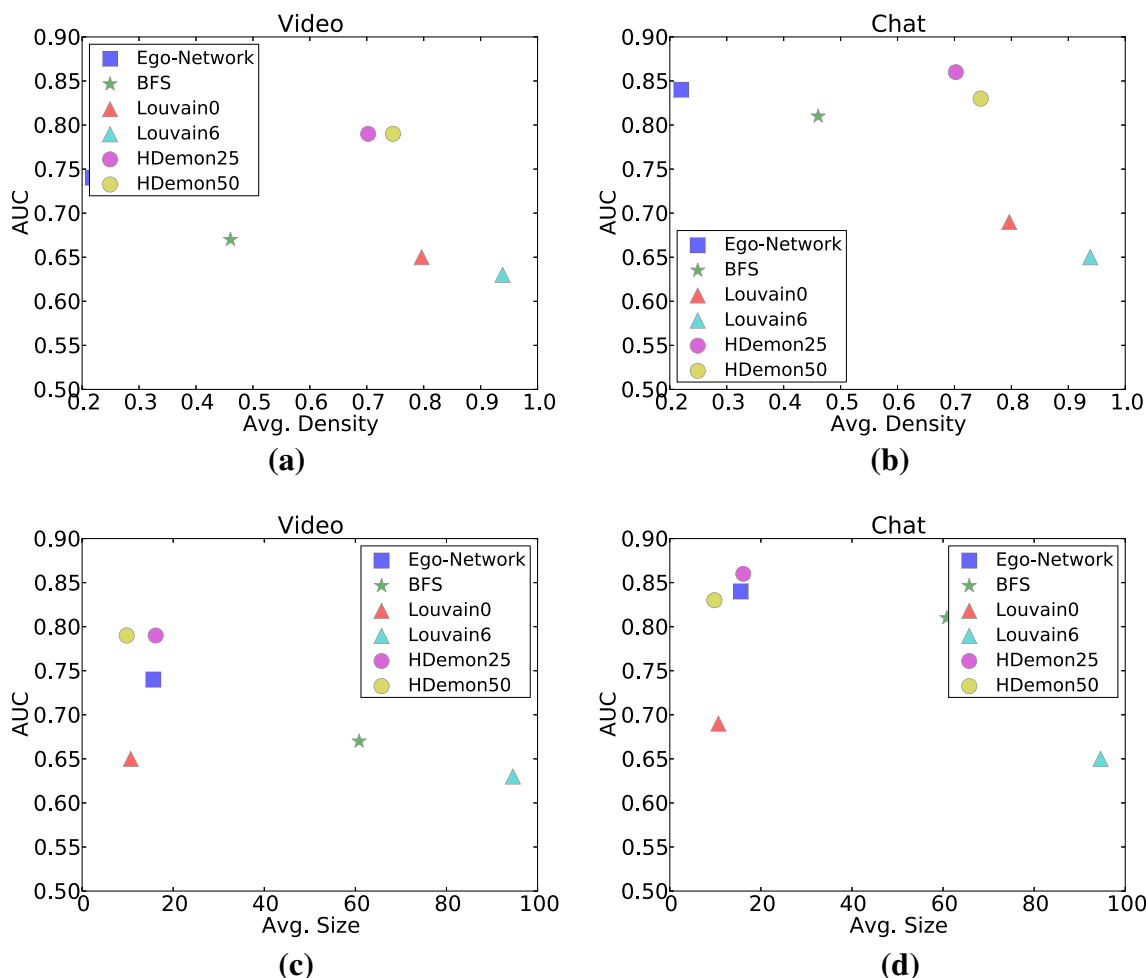
4.1.2 Unbalanced scenario

We address also an unbalanced scenario where we use the 75th percentile for the low engagement class, which thus contains the 75 % of the observations, and put the remaining 25 % of the observations in the high engagement class. Table 4 describes the results produced by the SGD methods in the unbalanced scenario, using the same features and community discovery approaches discussed before. The baseline method for the unbalanced scenario is the majority classifier: it reaches an AUC of 0.75 by assigning each item to the majority class (the low engagement class). We observe that, regardless the community set used, the SGD method (as well as random forest) is not able to improve significantly the baseline classifier for video. Conversely the results obtained for the chat feature by SGD outperform the baseline when we adopt HDEMON, EGO-NETWORKS and BFS community sets, reaching an AUC of 0.83.

In order to provide additional insights into the models built with the adoption of the different CD algorithms, we also compute the precision and recall measures with respect to the minority class (see Table 5). Looking at these measures enables us to understand which is the advantage in using SGD to identify correctly instances of the less

Fig. 4 Skype: AUC versus avg. density and AUC versus avg. size for video and chat in the balanced scenario. **a** AUC versus density: video. **b** AUC versus density: chat. **c** AUC versus size: video. **d** AUC versus size: chat

predictable class. Moreover, we can observe how choosing the 75th percentile led to a very difficult classification setup: the instances belonging to the minority class often represent outliers having very few examples from which the classifier can learn the model. Here the baseline is the minority classifier which reaches a precision of 25 % by assigning each community item to the minority class (the high engagement one). We observe that the SGD method outperforms the baseline classifier on all the community sets (reaching values in the range [.33, .57]). HDEMON and EGO-NETWORKS are the community sets which led to the best precision, on the video features and the chat feature, respectively.

In order to measure the effectiveness of SGD we report the lift chart which shows the ratio between the results obtained with the built model and the ones obtained by a random classifier. The charts in Fig. 5 are visual aids for measuring SGD's performance on the community sets: the greater the area between the lift curve and the baseline, the better the model. We observe that HDEMON performs better

than the competitors for the video features. For the chat features, the community sets produced by the three naive algorithm win against the other two CD algorithms. For all the three activity features, LOUVAIN reaches the worst performance, as in the balanced scenario.

As done for the balanced scenario in Fig. 3 we report the features having weight greater than 0.2 or lower than −0.2. In contrast with the results presented in the previous section, where topological features alway show the higher relative importance for the classification process, in this scenario we observe how *community formation* and *geographical* features are the ones which ensure greater descriptive power. As previously observed the minority class identified by a 75th percentile split is mostly composed by particular, rare, community instances. This obviously affects the relative importance of temporal and geographical information: the results suggest that the more a community is active the more significative are its geographical and temporal bounds. Finally in Fig. 6 we show the relationships between the average community size, the

Table 4 Skype: AUC and accuracy (within brackets) produced by the SGD method in the unbalanced scenario, for the video and chat features

Algorithm	Lv.	Scores
Video: AUC and accuracy		
HDemon25	1	**.76** (.68)
HDemon50	0	.73 (.65)
Louvain	0	.64 (.59)
Louvain	6	.61 (.58)
Ego-Nets	–	.71 (.63)
BFS	–	.68 (.61)
Baseline	–	.75
Chat: AUC and accuracy		
HDemon25	2	.82 (.78)
HDemon50	3	.80 (.76)
Louvain	0	.68 (.70)
Louvain	6	.67 (.66)
Ego-Nets	–	**.83** (.79)
BFS	–	.82 (.77)
Baseline	–	.75

In bold the best model. The baseline method is the majority classifier, which reaches an AUC of 0.75 by assigning each item to the majority class (the low engagement class)

Table 5 Skype: precision and recall (within brackets) produced by the SGD model for the video and chat features in the unbalanced scenario

Algorithm	Lv.	Scores
Video: precision–recall		
HDemon25	2	**.42** (.72)
HDemon50	1	.39 (.70)
Louvain	0	.33 (.69)
Louvain	6	.33 (.67)
Ego-Nets	–	.37 (.68)
BFS	–	.35 (.71)
Baseline	–	.25
Chat: precision–recall		
HDemon25	2	.54 (.69)
HDemon50	3	.50 (.67)
Louvain	0	.40 (.41)
Louvain	6	.44 (.33)
Ego-Nets	–	**.57** (.68)
BFS	–	.52 (.71)
Baseline	–	.25

In bold the best model. Having used the 75th percentile to discriminate the class labels the precision baseline w.r.t. the positive class is .25

average community density and the AUC value produced by the SGD method on the community sets which reach the best performances in the unbalanced scenario. We can observe how, in this settings, the algorithms producing communities with small average sizes and high density are the ones that assure the construction of SGD models reaching higher AUC. In particular HDemon in both its instantiation outperforms the other approaches.

4.1.3 Skype community characterization

From our analysis a well-defined trend emerges: among the compared methodologies, in both the balanced and unbalanced scenarios, HDemon is the best in bounding homophily producing communities that guarantee useful insights into the product engagement level. For this reason starting from the communities extracted by such bottom-up overlapping approach we computed the Pearson correlation for all the defined features against the final class label (high/low engagement). As shown in Fig. 7a when splitting the video engagement using the 50th percentile we are able to identify as highly active communities the ones having high country entropy E_s as well as high geographical distance among its users $dist_{avg}$ and whose formation is recent (i.e., whose first user has joined the network recently, T_f, as well as the last one, $IT_{l,f}$.). Moreover, video active communities tends to be composed by users having on average low degree as shown by deg_{avg}^{all} and deg_{max}^{C}. Conversely looking at Fig. 7b we can notice that communities which exhibit high chat engagement can be described by persistent structures (i.e., social groups for which the inter-arrival time $IT_{l,f}$ from the first to the last user is high), composed by users showing almost the same connectivity (in particular having high degree) and sparse social connections (low clustering coefficient CC, low density D and high radius). Moreover, we calculate the same correlations for the 75th percentile split: in contrast with the new results for the chat engagement (Fig. 7d) which do not differ significantly from the ones discussed for the balanced scenario, in this settings the highly active video communities show new peculiarities. In Fig. 7c we observe how the level of engagement inversely correlates with the community radius (and diameter) and directly correlates with density. This variation describes highly active video communities as a specific and homogeneous subclass composed by small and dense network structures composed by users who live in different countries (high geographical entropy E_s).

4.2 LastFM: service engagement

For the LastFM scenario we want to understand if the topological features of the social network can explain whether a community is predictive of the engagement into the service, measured by the total number of listenings of users into the community. To do that we transform the
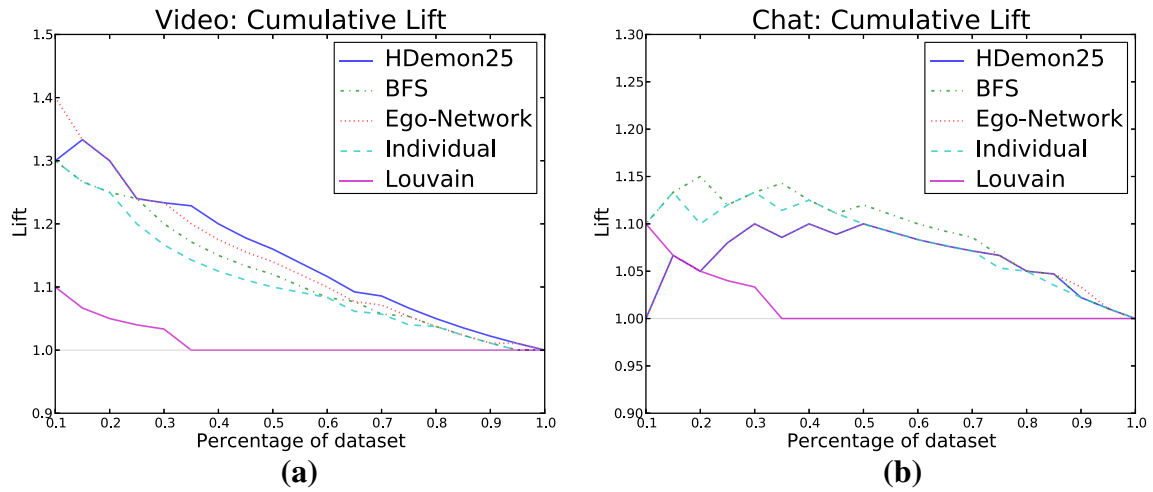
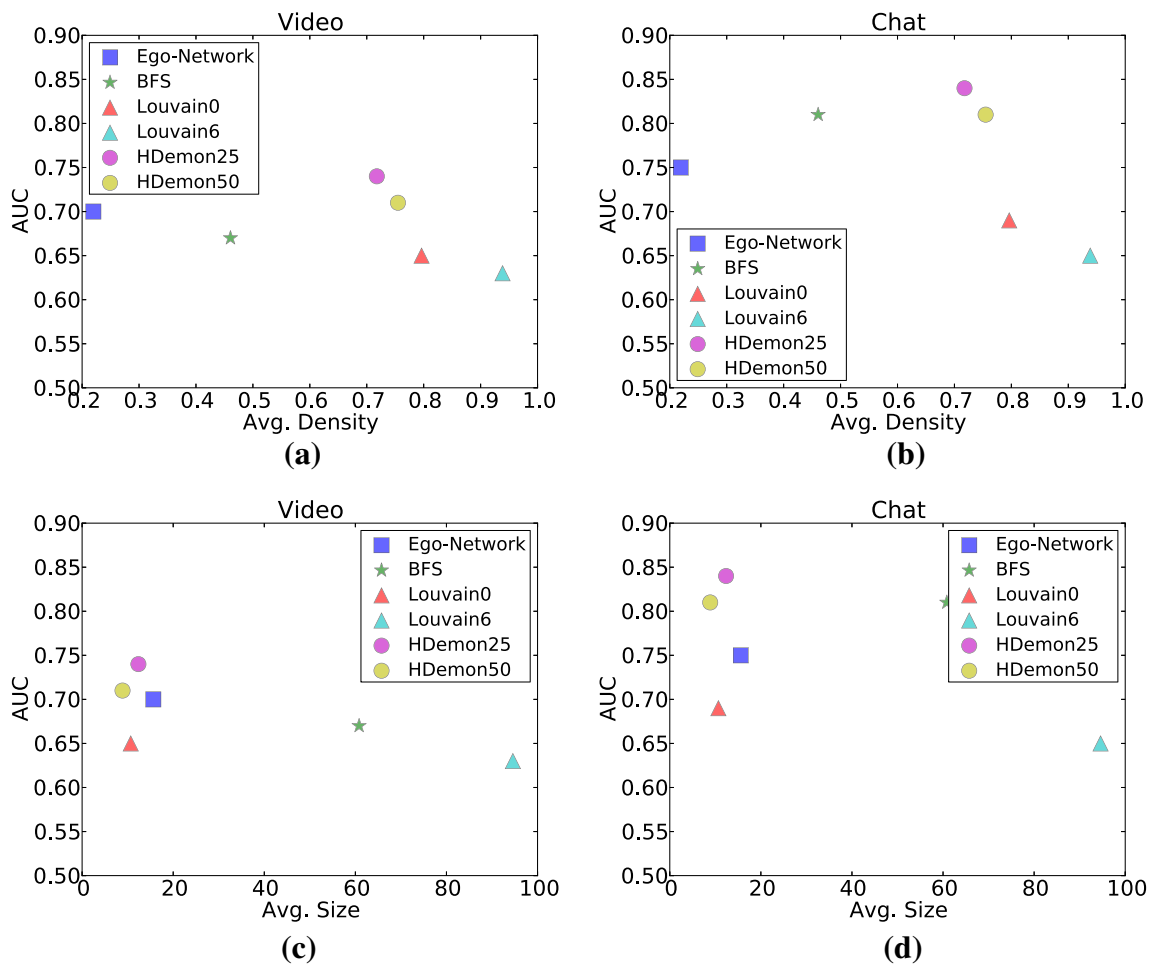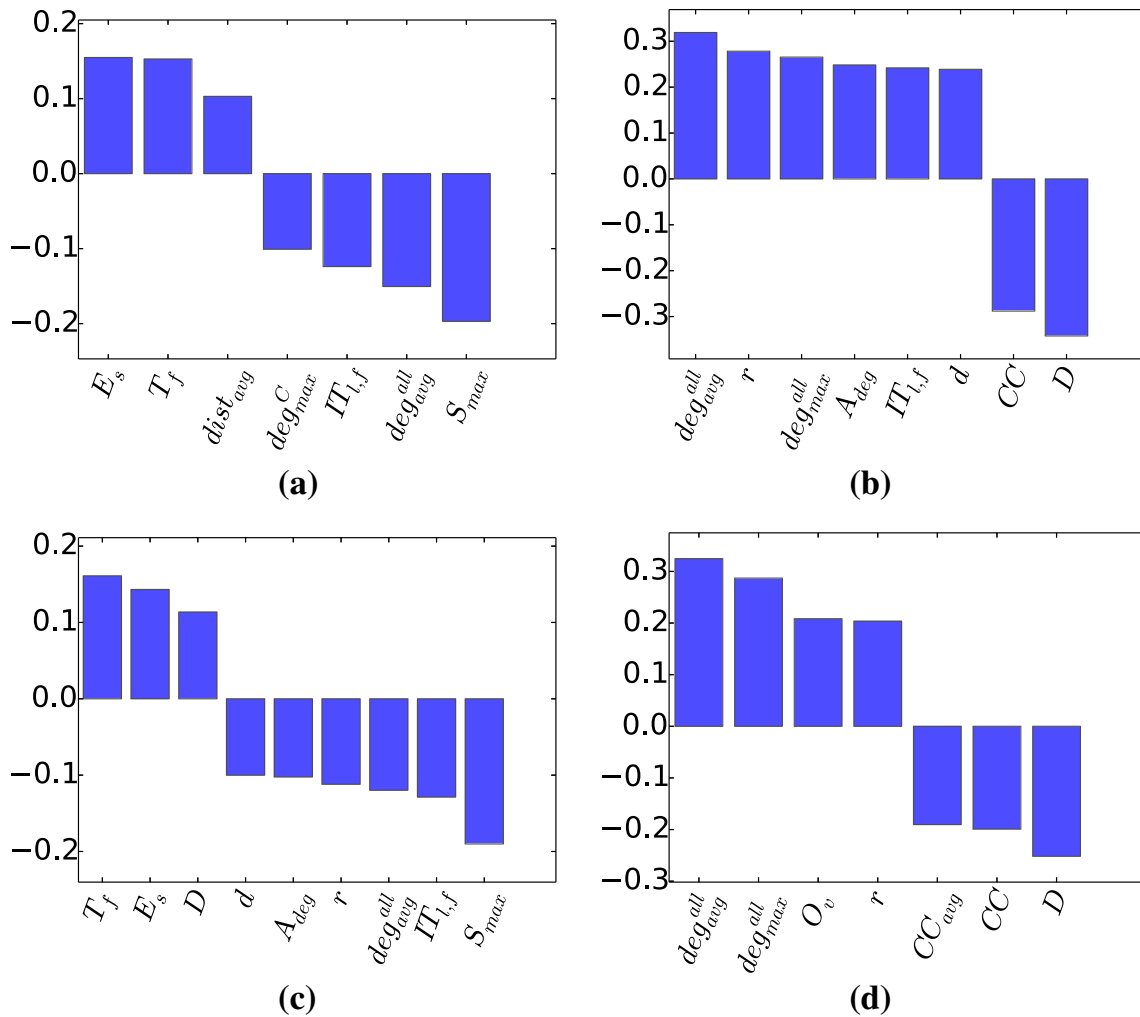**Fig. 5** Skype: lift plot for video and chat in the unbalanced scenario. **a** Video. **b** Chat

Fig. 6 Skype: AUC versus avg. density and AUC versus avg. size for video and chat in the unbalanced scenario. **a** AUC versus density: video. **b** AUC versus density: chat. **c** AUC versus size: video. **d** AUC versus size: chat

Fig. 7 Skype: most relevant Pearson correlations between community feature values and target class (high/low activity) for HDEMON. In **a, b** are shown the indexes for the balanced class scenario while in **c, d** for the 75th percentile split

Table 6 LastFM: AUC and accuracy (within brackets) produced by the best classifier in the balanced scenario, for the average total listenings feature

Algorithm	Scores	Classifier
LastFM: AUC and accuracy		
DEMON	.59 (.63)	Logistic regression
LOUVAIN	**.71** (.72)	Decision tree
EGO-NETS	.55 (.57)	Logistic regression

In bold the best model

problem into a binary classification task by assigning each community to one of the two classes: low volume of listenings or high volume of listenings. As for the Skype network, we address two different scenarios: (1) a balanced class scenario where the two classes have the same percentage of population (50th percentile split) and (2) an unbalanced class scenario (75th percentile split) where we consider an uneven class distribution.

4.2.1 Balanced scenario

The results reported in Table 6 highlight how, in contrast with Skype, LOUVAIN produces the best performance in predicting the volume of listenings (both in AUC and accuracy). This trend is also evident from Fig. 8: LOUVAIN shows lower average density and lower average size than the other algorithms, albeit obtaining the highest AUC. The EGO-NETS approach produces the worst performance highlighting how, in a balanced scenario, the community-based approach improves the prediction of the engagement.

4.2.2 Unbalanced scenario

In the unbalanced scenario the low volume of listenings class is the 75 % of the dataset. Tables 7 and 8 show two main results. On the one hand, HDEMON produces the best performance reaching an AUC = .78 (Table 7), a considerable improvement with respect to the baseline classifier
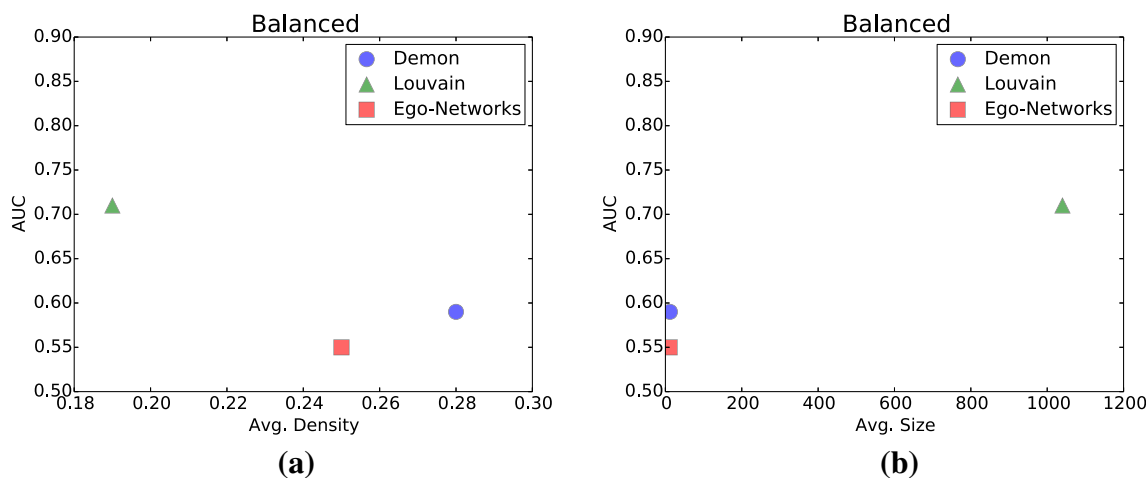
Fig. 8 LastFM: AUC versus avg. density and AUC versus avg. size in the balanced scenario. **a** AUC versus density: LastFM. **b** AUC versus size: LastFM

Table 7 LastFM: AUC and accuracy (within brackets) produced by the best classifier in the unbalanced scenario, for the average total listening feature

Algorithm	Scores	Classifier
LastFM: AUC and accuracy		
DEMON	**.60** (.78)	Logistic regression
LOUVAIN	.55 (.36)	Logistic regression
EGO-NETS	.55 (.83)	Random forest
Baseline	.25 (.25)	–

In bold the best model. The baseline method is the majority classifier, which reaches an AUC of 0.75 by assigning each item to the majority class (the low engagement class)

Table 8 LastFM: precision and recall (within brackets) produced by the best classifier for the average total listenings feature in the unbalanced scenario

Algorithm	Scores	Classifier
LastFM: precision–recall		
DEMON	.78 (.03)	Logistic regression
LOUVAIN	.33 (.30)	Decision tree
EGO-NETS	**.83** (.004)	Random forest
Baseline	.25 (1.0)	–

In bold the best model. Having used the 75th percentile to discriminate the class labels the precision baseline w.r.t. the positive class is .25

(.25). Figure 9 shows that HDEMON communities are the ones whose topological attributes better discriminate among the high volume and low volume listenings classes. On the other hand, the EGO-NETS algorithm produces the best precision on the minority class (Table 8). In any case all the algorithms outperform the baseline precision on the minority class (0.25), even though they show a rather low recall (while the baseline by definition has recall = 1).

4.3 Google+: community homogeneity

In this scenario we investigate the ability of topological features in explaining whether a community is composed by users having a homogeneous level of education. As done before, we see the problem as a binary classification task, i.e., each community is assigned to one of the two classes: (1) homogeneous or (2) heterogeneous education level. The target feature is built computing the node label entropy e_i for each community c_i: if $e_i \rightarrow 0$ community users have the same education level, conversely if $e_i \rightarrow 1$ they show heterogeneous education levels. The chosen target feature distributes almost equally on all the partitions made, following a normal distribution. We address two different scenarios: (1) a balanced class scenario where the two classes have the same percentage of population (50th percentile split) and (2) an unbalanced class scenario (75th percentile split), where we consider an uneven class assignment (rising the threshold level for homogeneous communities).

4.3.1 Balanced scenario

As done for LastFM, since the dataset has moderate size we applied an ensemble of classification approaches and report the results obtained by the best performer. The results reported in Table 9 highlight how, contrarily to what observed on Skype, LOUVAIN guarantees the best performances (both in AUC and accuracy). This trend is evident in Fig. 10: LOUVAIN seems to better capture the degree of homophily because—due to the scale problem that affects modularity-based approaches—it outputs huge
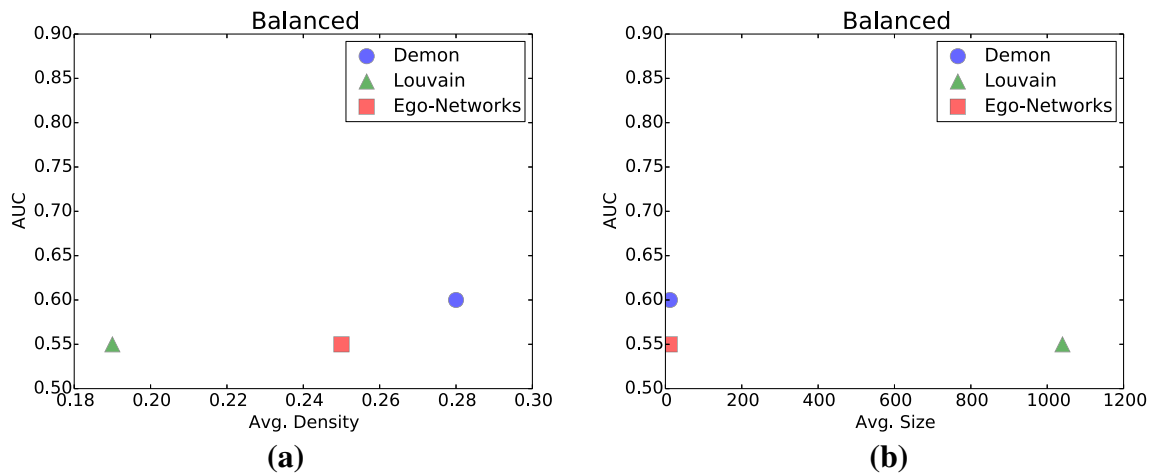
Fig. 9 LastFM: AUC versus avg. density and AUC versus avg. size in the unbalanced scenario. **a** AUC versus density: LastFM. **b** AUC versus size: LastFM

Table 9 Google+: AUC and accuracy (within brackets) produced by the best classifier (SGD) applied to the Google+ topological features in the balanced scenario

Algorithm	Scores	Classifier
Google+: AUC and accuracy		
DEMON	.67 (.71)	SGD
LOUVAIN	**.74** (.84)	SGD
EGO-NETS	.61 (.65)	SGD
Baseline	.50 (.50)	–

communities (whose entropy tends to 1) and tiny communities (whose entropy tends to 0).

The reduced quality of prediction obtained by HDEMON and EGO-NETWORK highlights the complexity of the problem: EGO-NETWORKS guarantee smaller and denser communities, but fail in recovering all the positive instances (low recall on the homogeneous class, $\simeq 0.41$); HDEMON reaches a higher recall but, due to the higher average sizes of the identified communities, lacks in precision ($\simeq 0.52$).

4.3.2 Unbalanced scenario

We applied the same strategy to address a more complex scenario: in this settings the homogeneous level of education is assigned only to communities having node label entropy in the range [0, 0.25]. We are searching for the most homogeneous communities.

Tables 10 and 11 show that the best classification is reached when the HDEMON communities are used. As expected LOUVAIN performances decrease while focusing on the minority class (which contains small- and medium-sized communities). From Table 11 we get a very clear picture on the complexity of the problem itself: all the
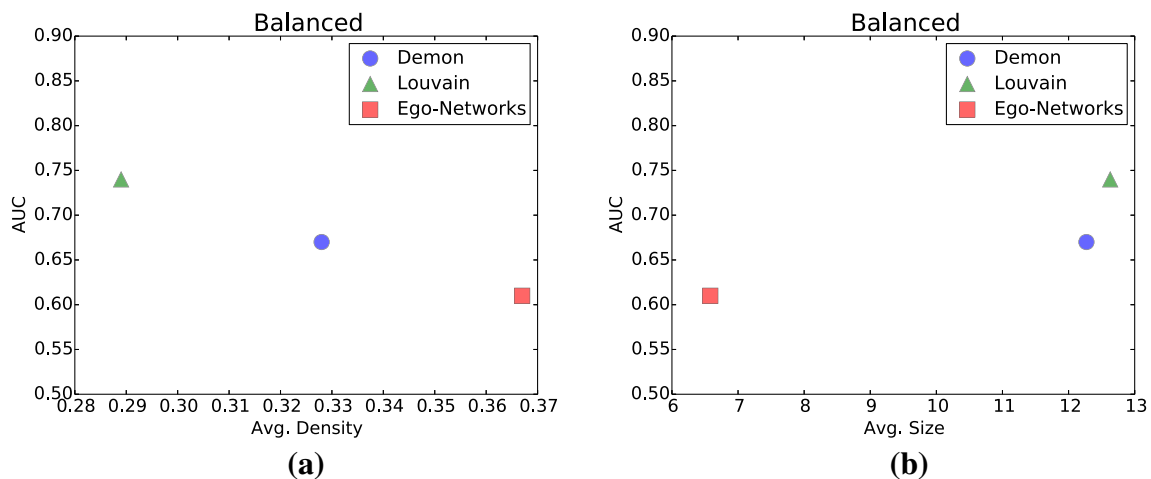
Fig. 10 Google+: AUC versus avg. density and AUC versus avg. size in the balanced scenario. **a** AUC versus density: Google+. **b** AUC versus size: Google+

Table 10 Google+: AUC and accuracy (within brackets) produced by the best classifier (decision tree) applied to the Google+ topological features in the unbalanced scenario

Algorithm	Scores	Classifier
Google+: AUC and accuracy		
DEMON	**.69** (.70)	Decision tree
LOUVAIN	.61 (.50)	Decision tree
EGO-NETS	.63 (.50)	Decision tree
Baseline	.75	–

In bold the best model

Table 11 Google+: precision and recall (within brackets) produced by the best classifier (decision tree) applied to the Google+ topological features in the unbalanced scenario

| Algorithm	Scores	Classifier
Google+: precision–recall		
DEMON	**.70** (.22)	Decision tree
LOUVAIN	.50 (.03)	Decision tree
EGO-NETS	.50 (.04)	Decision tree
Baseline	.25	–

In bold the best model

proposed community discovery algorithms outperform the baseline precision on the minority class; however, their recall is quite low (while the baseline, by definition has recall = 1). Again Fig. 11 shows that HDEMON communities are the best in discriminating among homogeneous and heterogeneous users education level.

5 Discussion

After the analysis of three different datasets we make some general observations on the obtained results. Two important aspects need to be evaluated while addressing the *homophilic network decomposition* problem on a given dataset:

- The social network semantic (i.e., which kind of relation is defined by edges? Are the links among nodes viable proxy for real social connections?);
- The nature of the target features.

In our applicative scenarios we instantiate the general problem on online scenarios having different peculiarities w.r.t. both these aspects. The Skype dataset, our primary playground, can be considered a trustable social proxy: each edge represents a connection among two users that know each other. Moreover, the usage information of video and chat, although individual, can be seen as a proxy for the communication among connected users. In such scenario we can assume that users within a community intrinsically cooperate to reach a certain level of activity w.r.t. a specific product/service. In LastFM, even if we are still analyzing a social structure, the target attribute relates to an average individual activity. While in Skype the usage of chat/video within a community is likely to involve all the users within the community, in LastFM the usage of the platform is defined by individual actions. Finally on Google+ the target regards personal information, which represents one of the reasons behind the presence of some network connections (i.e., if they studied together) but that it is not necessarily the glue that keeps communities together.
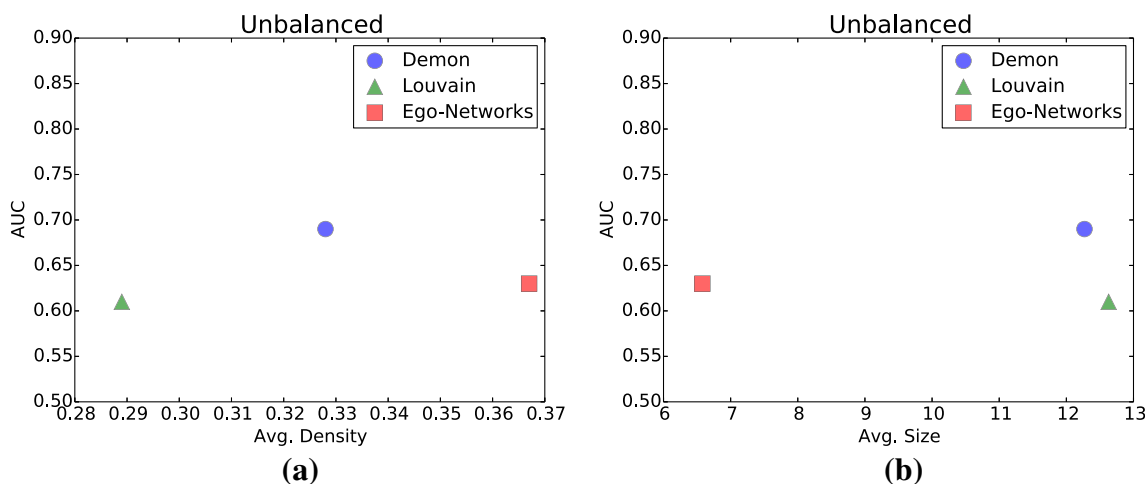
Fig. 11 Google+: AUC versus avg. density and AUC versus avg. size in the unbalanced scenario. **a** AUC versus density: Google+. **b** AUC versus size: Google+

These differences among the considered scenarios are the main reasons of the different outcomes the proposed approach produces. For example, while in Skype the community algorithms producing small and dense communities (e.g., HDEMON) guarantee the best solutions to our problem, in LastFM and Google+ the modularity-based algorithms tend to outperform the others. These results that at first sight can appear conflicting are instead a clear evidence that the network semantics and the definition of target features have a great impact on the problem solution. Moreover, as shown in Sect. 3 we exploit geographical features in order to improve the level of homophily across the nodes within communities: we include such information both implicitly—in LastFM and Google+ the selected users all have the same nationality—and where available explicitly—as for Skype. In the Skype scenario we observe that geographical proximity entropy information can be used to explain differently each specific target feature to predict: even though in social networks it is easy to observe several homophilic phenomena on top of the same structure, it is possible to identify different partitions able to guarantee high homogeneity w.r.t. specific attributes.

6 Related works

In this work we address the problem of predicting the degree of homophily of communities from their network topology. Homophily (McPherson et al. 2001) is a widely studied property that permeates different social networks: in recent studies, homophily has been leveraged to boost classical graph mining tasks such as link prediction (Elkabani and Khachfeh 2015; Yuan et al. 2014; Rossetti et al. 2015) and community discovery (Zardi et al. 2014), to build recommendation systems (Carullo et al. 2015; Zhao et al. 2014; Wang et al. 2013) and to study diffusion of (mis)information (Bessi et al. 2015).

6.1 Activity prediction and social targeting

In the Skype and LastFM scenarios we define clear examples of how the general issue we defined can be instantiated in very specific contexts. User/product engagement analysis is one of the most valuable fields of research for companies that needs to promote their services on targeted audiences: in recent years, many works addressed the issue of predicting users' future activities based on their past social behavior, thanks to the fertile ground provided by social media like Facebook and Twitter. For example, Zhu et al. (2013) conduct experiments on the social media Renren using a social customer relationship management (Social CRM) model, obtaining superior performance when compared with traditional supervised learning methods. Other works focus in particular on the prediction of churn, i.e., the loss of customers. Oentaryo et al. (2012) propose a churn prediction approach based on collective classification (CC), evaluating it using real data provided by the myGamma social networking site. They demonstrate that using CC on structural network features produces better predictions than conventional classification on user profile features. Richter et al. (2010) analyze a large call graph to predict the churn rate of its customers. They defines the churn probability of a customer as a function of its local influence with immediate social circle and the churn probability of the entire social circle as obtained from a predictive model.

A different category of works focus on online advertisement and market targeting on social networks. Bhatt et al. (2010) address the problem of online advertising by analyzing user behavior and social connectivity on online social networks. Studying the adoption of a paid product by members of the Instant Messenger (IM) network, they first observe that the adoption is more likely if the product has been widely adopted by the individual's friends. They then build predictive models to identify individuals most suited for marketing campaigns, showing that predictive models for direct and social neighborhood marketing outperform several widely accepted marketing heuristics. Domingos and Richardson (2001) propose to evaluate a user's network value in addition to their intrinsic value and its effectiveness in viral marketing, while Hartline et al. (2008) propose a strategy wherein a carefully chosen set of users is influenced with free distribution of the product and the remaining buyers are exploited for revenue maximization. Authors of Bagherjeiran and Parekh (2008) present a machine learning approach which combines user behavioral features and social features to estimate the probability that a user to click on a display ad.

6.2 Community detection in social networks

One challenging problem in network science is the discovery of communities within the structure of complex networks. Two surveys by Fortunato (2010) and Coscia et al. (2012) explore the most popular community detection techniques and try to classify algorithms given the typology of the extracted communities. One of the most adopted definitions of community is based on the modularity concept (Newman and Girvan 2004; Clauset et al. 2004), a quality function of a partition which scores high values for partitions whose internal cluster density is higher than the external density. The seminal algorithm proposed by Girvan and Newman (2002) and Newman and Girvan (2004) iteratively removes links based on the value of their betweenness, i.e., the number of shortest paths that pass through the link. The procedure of link removal ends when

the modularity of the resulting partition reaches a maximum. The method introduced by Clauset et al. (2004) is essentially a fast implementation of a previous technique proposed by Newman and Girvan (2004). A fast and efficient greedy algorithm, Louvain, has been successfully applied to the analysis of huge subset of the WWW (Blondel et al. 2008). Modularity is not the only key concept that has been used for community detection: an alternative approach is the application of information theory techniques, as for example in Infomap (Rosvall and Bergstrom 2008). An interesting property for community discovery is the ability to detect overlapping substructures, allowing nodes to be part of more than one community. A wide set of algorithms are developed over this property, such as cFinder (Palla et al. 2005) and Demon (Coscia et al. 2014).

7 Conclusions

In this work we formulated the problem of *homophilic network decomposition*. After the formulation of the general problem we instantiated it on different scenarios: user/service engagement analysis and attribute homogeneity evaluation. We first produced several community sets from the global Skype network by applying different community detection algorithms on the data. We then extracted from each community topological, geographical and temporal features and learned classification models to predict the level of usage for the video and chat products (Skype), the average level of listening of users (LastFM) and the homogeneity of the education level in a community (Google+). On the Skype network, our results showed that algorithms producing overlapping micro-communities like HDemon reach the best performances. Conversely modularity-based approaches like Louvain do not guarantee good performance and are often outperformed by naive algorithms such as Ego-Nets and BFS. Subsequently we applied the same analytical framework to LastFM and Google+. In contrast with the results observed on Skype, in these scenarios Louvain is the best approach in capturing homophilic behavior. These counterintuitive results are due to the different nature of the analyzed services and target features: while the user engagement in Skype is strictly related to the users within a community (and the final aim of the network itself), the service engagement and education level are only averages of individual peculiarities (thus more difficult to relate to community structures).

Our results could be further improved by two properties which are not present in the analyzed datasets: the strength of the ties between the users and the dynamics of user profiles and network links. On one side, tie strength quantifies the degree of interaction between two individuals, allowing to understand at what extent the level of interactions inside a community is a proxy for users homogeneity w.r.t. a specific feature. On the other side, temporal information about the appearance/vanishing of links as well as the geographical location of users allows us to investigate how network and community structures change in time, thus avoiding over/underestimation of the real sociality as observed in a static network scenario.

References

Bagherjeiran A , Parekh R (2008) Combining behavioral and social network data for online advertising. In: ICDM workshops

Bessi A, Petroni F, Vicario MD, Zollo F, Anagnostopoulos A, Scala A, Caldarelli G, Quattrociocchi W (2015) Viral misinformation: the role of homophily and polarization. In: Proceedings of the 24th international conference on world wide web companion, WWW 2015, Florence, Italy, May 18–22, 2015—companion volume

Bhatt R, Chaoji V, Parekh R (2010) Predicting product adoption in large-scale social networks. In: CIKM

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008

Carullo G, Castiglione A, Santis AD, Palmieri F (2015) A triadic closure and homophily-based recommendation system for online social networks. World Wide Web 18(6): 1579–1601 (Online). doi:10.1007/s11280-015-0333-5

Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Rev E Phys

Coscia M, Giannotti F, Pedreschi D (2012) A classification for community discovery methods in complex networks. In: CoRR

Coscia M, Rossetti G, Giannotti F, Pedreschi D (2014) Uncovering hierarchical and overlapping communities with a local-first approach. In: TKDD

Domingos P, Richardson M (2001) Mining the network value of customers. In: SIGKDD

Elkabani I, Khachfeh RAA (2015) Homophily-based link prediction in the facebook online social network: a rough sets approach. J Intell

Syst 24(4):491–503 (Online). http://www.degruyter.com/view/j/jisys.2015.24.issue-4/jisys-2014-0031/jisys-2014-0031.xml

Fortunato S (2010) Community detection in graphs. Phys Rep 486(3):75–174

Fortunato S, Barthélemy M (2007) Resolution limit in community detection. IN: PNAS

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. In: PNAS

Gong NZ, Xu W, Huang L, Mittal P, Stefanov E, Sekar V, Song D (2012) Evolution of social-attribute networks: measurements, modeling, and implications using google+. CoRR abs/1209.0835 (Online). http://arxiv.org/abs/1209.0835

Hartline JD, Mirrokni VS, Sundararajan M (2008) Optimal marketing strategies over social networks. In: WWW

Himelboim I, McCreery S, Smith M (2013) Birds of a feather tweet together: integrating network and content analyses to examine cross-ideology exposure on twitter. J Comput Med Commun 18(2):40–60

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. In: Annual review of sociology

Newman MEJ (2003) Mixing patterns in networks. Phys Rev E 67:026126

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113

Oentaryo RJ, Lim E-P, Lo D, Zhu F, Prasetyo PK (2012) Collective churn prediction in social network. In: ASONAM

Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nat 435(7043):814–818

Richter Y, Yom-Tov E, Slonim N (2010) Predicting customer churn in mobile networks through analysis of social groups. In: SDM

Rossetti G, Guidotti R, Pennacchioli D, Pedreschi D, Giannotti F (2015) Interaction prediction in dynamic networks exploiting community discovery. In: International conference on advances in social network analysis and mining, IEEE, pp 553–558 (Online). http://dl.acm.org/citation.cfm?doid=2808797.2809401

Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. In: PNAS

Tsuruoka Y, Tsujii J, Ananiadou S (2009) Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In: ACL/IJCNLP

Wang Y, Zang H, Faloutsos M (2013) Inferring cellular user demographic information using homophily on call graphs. In: 2013 Proceedings IEEE INFOCOM workshops, Turin, Italy, 14–19 Apr 2013, pp 211–216 (Online). doi:10.1109/INFCOMW.2013.6562897

Watts D, Strogatz S (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442

Yuan G, Murukannaiah PK, Zhang Z, Singh MP (2014) Exploiting sentiment homophily for link prediction. In: Eighth ACM conference on recommender systems, RecSys '14, Foster City, Silicon Valley, CA, 06–10 Oct 2014, pp 17–24 (Online). doi:10.1145/2645710.2645734

Zardi H, Romdhane LB, Guessoum Z (2014) A multi-agent homophily-based approach for community detection in social networks. In: 26th IEEE international conference on tools with artificial intelligence, ICTAI 2014, Limassol, Cyprus, 10-12 Nov 2014, pp 501–505 (Online). doi:10.1109/ICTAI.2014.81

Zhang T (2004) Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: ICML

Zhao T, Hu J, He P, Fan H, Lyu MR, King I (2014) Exploiting homophily-based implicit social network to improve recommendation performance. In: 2014 International joint conference on neural networks, IJCNN 2014, Beijing, China, 6–11 July 2014, pp 2539–2547. (Online). doi:10.1109/IJCNN.2014.6889743

Zhu Y, Zhong E, Pan SJ, Wang X, Zhou MQY (2013) Predicting user activity level in social networks. In: CIKM