Community-centric analysis of user engagement in Skype social network

Giulio Rossetti*†, Luca Pappalardo*†, Riivo Kikas‡, Dino Pedreschi*, Fosca Giannotti† and Marlon Dumas‡
*University of Pisa, Italy Email: {giulio.rossetti,lpappalardo,pedre}@di.unipi.it
†ISTI-CNR, Pisa Italy Email: {name.surname}@isti.cnr.it
‡Unversity of Tartu, Estonia Email: {riivokik, marlon.dumas}@ut.ee

***Abstract*—Traditional approaches to user engagement analysis focus on individual users. In this paper we address user engagement analysis at the level of groups of users (social communities). From the entire Skype social network we extract communities by means of representative community detection methods each one providing node partitions having their own peculiarities. We then examine user engagement in the extracted communities putting into evidence clear relations between topological and geographic features of communities and their mean user engagement. In particular we show that user engagement can be to a great extent predicted from such features. Moreover, from the analysis it clearly emerges that the choice of community definition and granularity deeply affect the predictive performance.**

I. 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 services are interested in exploiting these Big Data to understand “user engagement”, i.e. the way individuals use the products provided. Traditional approaches of predictive analytics focus on *individuals*: they try to describe and predict the level of engagement of a single individual, with the purpose of suggesting proper products/services and favoring the diffusion of the system over a larger population. Focusing on individuals, however, introduces many challenging issues, i.e., the amount of individuals to process is enormous, and hence hardly manageable. Addressing each single individual is also in many cases redundant, since neighbors in networks tend to behave in a similar way showing a certain degree of homophily [11], [10] and inevitably causes the underestimation of the surrounding social context. It is hence fundamental to widen the analysis spectrum to incorporate social surrounding of users in order to capture the homophily which characterize real social networks.

We propose to move the focus from individuals to groups analyzing and describing the engagement of *social communities*. 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 more complex objects from which we can extract a wide set of features for the analysis. To approach this problem, we extract social communities from the global Skype network and compute relevant structural and geographical features from each one of them. We then build a classifier to predict how much within a social community are used the video and instant messaging products provided by Skype. We find that group-centric approaches outperforms user-centric ones when we use algorithms producing overlapping micro-communities. In contrast modularity-based algorithms are worse than the ones of classical user-centric strategies. Hence, we show how the choice of a proper community detection algorithm is crucial to reach high performances in the engagement prediction.

II. Related Works

a) Activity prediction and social targeting: In recent years, many works addressed the issue of predicting users’ future activities based on their past social behavior. Zhu et al. [20] conduct experiments on the social media Renren using a Social Customer Relationship Management 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. [14] 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. [16] 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. [2] addresses 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 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. [7] propose to evaluate a user’s network value in addition to their intrinsic value and its effectiveness in viral marketing, while [9] propose

ASONAM '15, August 25-28, 2015, Paris, France

DOI: http://dx.doi.org/10.1145/2808797.2809384

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 [1] present a machine learning approach which combines user behavioral features and social features to estimate the probability that a user clicks on a display ad.

b) Community detection in social networks: One critical task of social network analysis involves the identification of groups and communities within complex social tissues. A survey [6] explore the most popular community detection techniques and try to classify algorithms given the typology of the extracted communities. One of the most adopted community definitions is based on the modularity concept [13], [4], a quality function of a partition which scores high values for partitions whose internal cluster density is higher than the external density. A fast and efficient modularity-based greedy algorithm, LOUVAIN, has been successfully applied to the analysis of huge subset of the WWW [3]. Moreover, 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 [17]. An interesting property for community discovery is the ability to detect overlapping sub-structures, allowing nodes to be part of more than one community. A wide set of algorithms are developed over this property, such as CFINDER [15], and DEMON [5].

III. MODEL CONSTRUCTION

Data: We analyze a dataset of users and connections in Skype as of October 2011. The dataset includes anonymized data of Skype users. Each user (identified by hashed ID) is associated with their account creation date and country and city of account creation. The dataset also includes connections between users. Connections are undirected: a connection exists between two users if and only if they belong to each other's contact list. Moreover, each connection is labeled with a timestamp corresponding to the contact request approval.

In addition to non-identifiable user profile data and network data, the dataset 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. 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. We analyze the most recent available snapshot of the network. Hence, we focus on the subset of users who used one of the two products, during at least two of the last three months covered in the dataset. Our analyses will be then executed on a filtered dataset composed by several tens of millions of users and connections.

Community Detection: The degree of *overlap* among communities is one of the properties that can be used to characterize community detection algorithms. Classical approaches produce crisp partition of the network, i.e. an individual can be involved in at most one community while overlapping ones considering the multidimensional nature of social networks allow individuals to belong to many different communities. To observe the impact overlap has on our analysis we use four different algorithms to extract social communities from the Skype network (in increasing degree of overlap): LOUVAIN, BFS, HDEMON and EGO-NETWORK.

LOUVAIN [3] is a scalable algorithm based on a greedy modularity approach. It 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 [8]. 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.

HDEMON [5] is based on a recursive hierarchical 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 algorithm has two parameters: (i) the minimum community size μ; and (ii) the minimum Jaccard ψ among meta-nodes to create an edge that connects them while building the community hierarchy. 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 produced the HDEMON25 community set, and $\psi = 0.5$ which produced the HDEMON50.

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 $\Gamma(u)$ identify 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. For each random user we extracted the corresponding ego network, filtering only unique ones.

The BFS algorithm extracts random 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, modeling community sizes. Starting from a root node, the algorithm explores other nodes performing a breadth first search and stopping when *csize* nodes are discovered.

Each algorithm, according to the specified parameters, produces different community sets when applied on the Skype dataset. In Table I we report for each community set and hierarchy level ($Lv.$) used in the following analysis: (i) the number of communities ($\#C$); (ii) the induced node coverage w.r.t. the whole graph; (iii) the average number of communities per node (σ, i.e. the mean degree of overlap); the average community size ($Avg.size$). LOUVAIN is a partitioning algorithm and guarantees the complete coverage of the nodes. HDEMON covers around 76% of the nodes because imposing the parameter $\mu = 3$ we exclude communities with two nodes only. BFS and EGO-NETWORK are executed on a 20% sample of the nodes, on which they cover the 90% and 69%

COMMUNITY STATISTICS

Algorithm	Lv.	$\#C$	coverage (%)	σ	Avg. size
HDEMON25	2	3.3e+07	76	13.2	27.9
HDEMON50	2	8.2e+07	76	10.3	8.9
LOUVAIN	0	8.7e+06	100	1.0	10.7
	6	9.8e+05	100	1.0	94.6
EGO-NETS	-	1.5e+07	69[1]	3.7	15.6
BFS	-	1.8e+07	90[1]	13.3	60.8

TABLE I: Characteristics of the community sets produced by the algorithms on the Skype dataset.

STRUCTURAL FEATURES

N	number of nodes
M	number of edges
D	density
CC	global clustering
CC_{avg}	average clustering
A_{deg}	degree assortativity
deg^{C}_{max}	max degree (community links)
deg^{C}_{avg}	avg degree (community links)
deg^{all}_{max}	max degree (all links)
deg^{all}_{avg}	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

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

GEOGRAPHIC 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 geographic distance
$dist_{max}$	max geographic distance

ACTIVITY FEATURES

Video	mean number of days of video
Chat	mean number of days of chat

TABLE II: Description of the features extracted from the communities.

respectively. For the LOUVAIN, we consider the hierarchical levels 0 and 6 only, which correspond to the first greedy iteration and the iteration having the maximum modularity.

A. Community Features

From the community sets produced by the four algorithms we extract a wide set of features, belonging to four main categories: *structural*, *geographical*, *formation* and *activity* features (see Table II). *Structural* features convey information about the topology of a social community. We analyze community size and density, clustering coefficient, diameter and radius as well as other relevant topological measures. Moreover we take into account as proxy for homophily the degree assortativity A_{deg} which indicates the preference for the nodes to attach to others that have the same degree [12]. 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 *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. *Geographic* features provide information about the geographic diversity of a community. 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. 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 to cities with more than 5,000 Skype users), we compute their geographic distance using the coordinates of the centers of the cities. Once computed all the available distances, we consider the average and the maximum geographic distances of each community. Finally, the *activity* features indicate the mean level of Skype activity performed by the community members. We extract two activity features: (i) *chat*, the mean number of days they used the instant messaging (chat); and (ii) *video*, the mean number of days they used the video conference. The distributions of the chat feature for HDEMON, BFS and EGO-NETWORKS follow a peaked distribution, while those of the chat feature (for LOUVAIN) and of the video feature (for all algorithms) follow an exponential distribution. In all cases, the separation between high-engagement and low-engagement communities is less clear for higher thresholds. For the video feature, the median ranges from 3 to 3.75 (across algorithms) while the 75th-percentile ranges from 6 to 7. For the chat feature, the median ranges from 5 to 5.9, while the 75th-percentile ranges from 13.9 to 15.4.

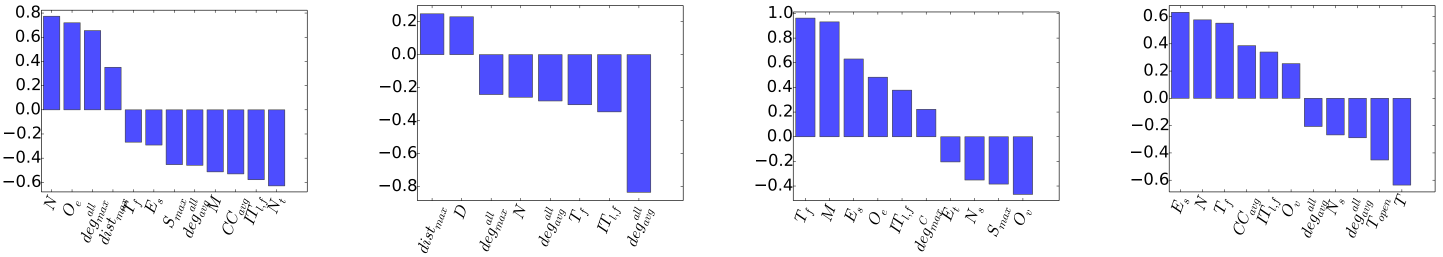
IV. MODEL EVALUATION

We use the 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: (i) a balanced class scenario where the two classes have the same percentage of population; and (ii) an unbalanced class scenario, where we consider an uneven population distribution.

Balanced scenario: In order to transform the video and chat activity features into discrete variables we partition the range of values through the median of their distribution. This produced, for each variable to predict, two equal-populated classes: (i) low engagement, ranging in the interval $[0, median]$; and (ii) high engagement, ranging in the interval $[median, 31]$.[2] To perform classification we use Stochastic Gradient Descent (SGD) and AUC (area under the ROC curve) to evaluate their performance. The overall accuracy is instead the proportion of true results (both true positives and true negatives) in the population. We learn the SGD classifier with logistic error function [18], [19] .We execute 5 iterations, performing data shuffling before each one of them, imposing the elastic-net penalty $\alpha = 0.0001$ and $l1\text{-}ratio = 0.05$. The adoption of elastic-net penalty results in some feature weights set to zero,

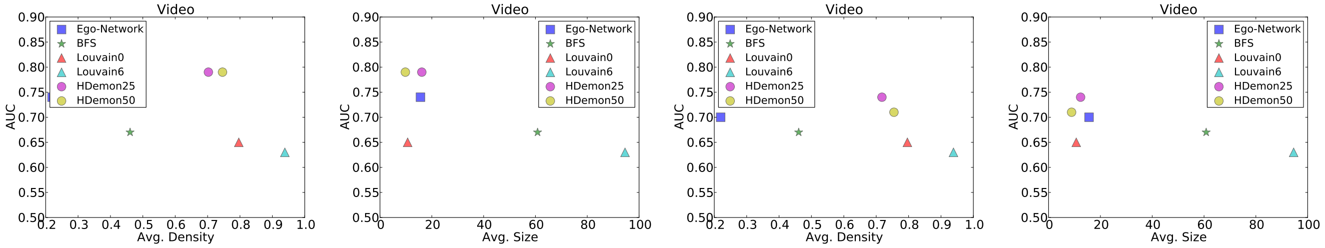
[1] For EGO-NETS and BFS the coverage is computed starting from a 20% sample of the total users.

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



(a) HDemon Chat (b) Louvain Chat (c) HDemon Video (d) Louvain Video

Fig. 1: Weights of the features produced by SGD method for HDemon and Louvain community sets, for the Chat feature in the balanced scenario (a-b) and Video feature in the unbalanced scenario (c-d).



(a) AUC vs. Density: Video (b) AUC vs. Size: Video (c) AUC vs. Density: Video (d) AUC vs. Size: Video

Fig. 2: AUC vs. Avg. Density and AUC vs. Avg. Size: Balanced scenario (a-b) Unbalanced scenario (c-d)

Video: AUC and Accuracy

Algorithm	Lv.	Scores
HDemon25	1	**.74** (.67)
HDemon50	0	.71 (.68)
Louvain	0	.65 (.60)
Louvain	6	.63 (.59)
Ego-Nets	-	.70 (.64)
BFS	-	.67 (.62)

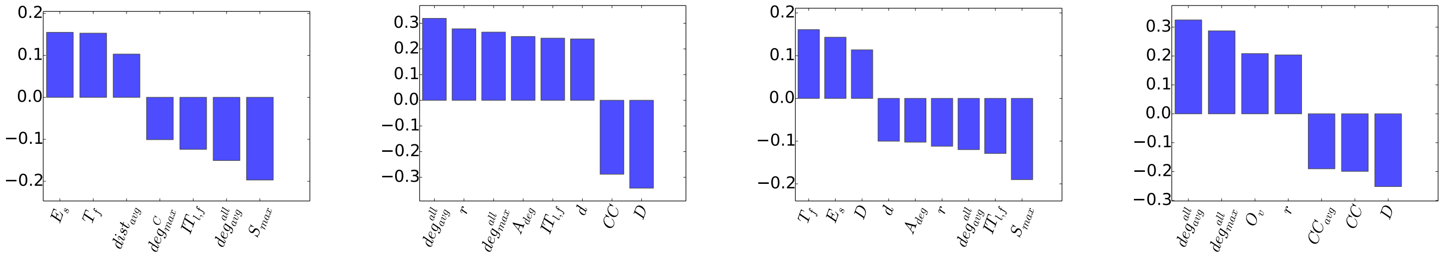
Chat: AUC and Accuracy

Algorithm	Lv.	Scores
HDemon25	2	**.84** (.77)
HDemon50	1	.81 (.73)
Louvain	0	.69 (.64)
Louvain	6	.65 (.60)
Ego-Nets	-	.75 (.75)
BFS	-	.81 (.72)

TABLE III: AUC and Accuracy (within brackets) in the balanced scenario, for Video and Chat.

thus eliminating less important features. We apply a five fold cross-validation for learning and testing. Table III 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 BFS and Ego-Networks. 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 [8]. As the level of the Louvain hierarchy increases, and hence the modularity increases, both the AUC and overall accuracy decrease. In the experiments, indeed, the first Louvain hierarchical level outperforms the last level, even though the latter has the highest modularity. Figure 1 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 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 Figure 1, d), confirming that a modularity approach produces communities with weak predictive power with respect to user engagement. Moreover, an interesting phenomena emerges: independently from the chosen community discovery approach, the most relevant class of features for the classification process is the *topological* class. In particular degree, density, community size and clustering related measures often appear among the most weighted features. Figures 2(a-b) 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 for the Video feature (Chat behave similarly). The best performance is obtained for the HDemon community sets, which constitute a compromise between the micro and the macro level of network granularity. 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 Louvain ones 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, reach the best performance in the problem of estimating user engagement.

Unbalanced scenario We address also an unbalanced scenario where we use the 75^{th} percentile to discriminate the low engagement class, which thus contains the 75% of the obser-



(a) Video median (b) Chat median (c) Video 75^{th} percentile (d) Chat 75^{th} percentile

Fig. 4: 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 75^{th} percentile split.

VIDEO: AUC AND ACCURACY

Algorithm	Lv.	Scores
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

Algorithm	Lv.	Scores
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

TABLE IV: AUC and Accuracy (within brackets) produced by the SGD method in the unbalanced scenario, for the Video and Chat features.

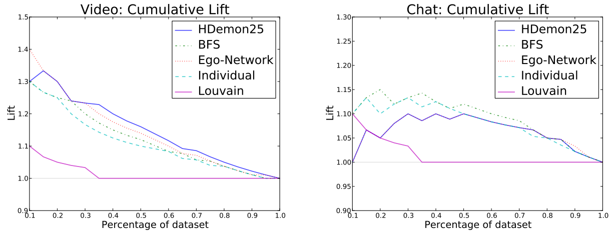
VIDEO: PRECISION - RECALL

Algorithm	Lv.	Scores
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

Algorithm	Lv.	Scores
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

TABLE V: Precision and Recall (within brackets) produced by the SGD model for the Video and Chat features in the unbalanced scenario.



(a) Video (b) Chat

Fig. 3: Unbalanced scenario: Lift plot for each product.

vations. Table IV 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 is not able to improve significantly the baseline classifier for Video. Conversely, the results obtained for the Chat feature by SGD outperforms the baseline when we adopt HDEMON, EGO-NETWORKS and BFS community sets, reaching an AUC of 0.83. In order to provide additional insights on the models built with the adoption of the different CD algorithms, we compute the precision and recall measures with respect to the minority class (see Table V). Looking at these measures enable us to understand which are the advantage in using SGD to identify correctly instances of the less predictable class. In this more challenging settings, 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 Figure 3 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 Figure 1(e-h) we report for each CD the features having weight greater than 0.2 or lower than −0.2. Conversely from the results presented in the previous section, where topological features always show the higher relative importance for the classification process, in this scenario we observe that *community formation* and *geographical* features have greater descriptive power. As previously observed the minority class identified by a 75^{th} percentile split is mostly composed by particular, rare, community instances affecting the relative importance of temporal and geographical informations: the results suggest that the more a community is active the more significative are its geographical and temporal bounds. Finally in Figure 2(c-d) we show 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 unbalanced scenario. We can observe how, in this settings, the algorithms granting communities having on average small 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.

V. COMMUNITY CHARACTERIZATION

From our analysis emerged a well defined trend: among the compared methodologies, HDEMON is able, both in balanced and unbalanced scenarios, to better bound homophily and thus to extract communities that guarantee useful insights on 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 Figure 4(a), when splitting the Video engagement using the 50^{th} percentile we are able to identify as highly active communities the ones having high country entropy E_s as well as high geographic 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 are composed by users having on average low degree as shown by deg^{all}_{avg} and deg^{C}_{max}. Conversely, looking at Figure 4(b) we 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 compute the same correlations for the 75^{th} percentile split: in contrast with the new results for the Chat engagement (Figure 4(d)) 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 Figure 4(c) we observe how the level of engagement negatively correlates with the community radius (and diameter) and positively correlates with the density. This variations describe highly active Video communities as a specific and homogeneous sub class composed by small and dense network structures composed by users who live in different countries (high geographical entropy E_s).

VI. CONCLUSIONS

In this work we addressed the issue of predicting user engagement in online social networks. In contrast with traditional user-centric approaches, we focus on social communities in order to exploit the inherent homophily characteristic of social networks. Our results show that, both in balanced and unbalanced classification scenarios, algorithms producing overlapping micro-communities like HDEMON reach the best performance. Conversely, modularity-based approach like LOUVAIN do not guarantee good performance and are outperformed by simple clustering strategies such as EGO-NETS and BFS. We also provide a description for low/high engaged communities identified by HDEMON through the analysis of the correlations between their activity level and the values of their features.

Acknowledgments. This research is supported by Microsoft/Skype and ERDF via the Software Technology and Applications Competence Centre (STACC) and partially funded by the European Community's H2020 Program under the funding scheme "FETPROACT-1-2014: Global Systems Science (GSS)", grant agreement #641191 CIMPLEX[3].

REFERENCES

[1] A. Bagherjeiran and R. Parekh, "Combining behavioral and social network data for online advertising." in *ICDM Workshops*, 2008.

[2] R. Bhatt, V. Chaoji, and R. Parekh, "Predicting product adoption in large-scale social networks." in *CIKM*, 2010.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[4] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, 2004.

[5] M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "Uncovering hierarchical and overlapping communities with a local-first approach." *TKDD*, 2014.

[6] M. Coscia, F. Giannotti, and D. Pedreschi, "A classification for community discovery methods in complex networks," *CoRR*, 2012.

[7] P. Domingos and M. Richardson, "Mining the network value of customers," in *SIGKDD*, 2001.

[8] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *PNAS*, 2007.

[9] J. D. Hartline, V. S. Mirrokni, and M. Sundararajan, "Optimal marketing strategies over social networks." in *WWW*, 2008.

[10] I. Himelboim, S. McCreery, and M. Smith, "Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter," *Journal of Computer-Mediated Communication*, 2013.

[11] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, 2001.

[12] M. E. J. Newman, "Mixing patterns in networks," *Phys. Rev. E*, vol. 67, p. 026126, 2003.

[13] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, 2004.

[14] R. J. Oentaryo, E.-P. Lim, D. Lo, F. Zhu, and P. K. Prasetyo, "Collective churn prediction in social network." in *ASONAM*, 2012.

[15] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, 2005.

[16] Y. Richter, E. Yom-Tov, and N. Slonim, "Predicting customer churn in mobile networks through analysis of social groups," in *SDM*, 2010.

[17] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *PNAS*, 2008.

[18] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty." in *ACL/IJCNLP*, 2009.

[19] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms." in *ICML*, 2004.

[20] Y. Zhu, E. Zhong, S. J. Pan, X. Wang, M. Zhou, and Q. Y. 0001, "Predicting user activity level in social networks." in *CIKM*, 2013.

[3] "Bringing CItizens, Models and Data together in Participatory, Interactive SociaL EXploratories'", https://www.cimplex-project.eu