Martin-Borregon et al. *EPJ Data Science* 2014, **2014**:8
http://www.epjdatascience.com/content/2014/1/8

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Characterization of online groups along space, time, and social dimensions

David Martin-Borregon[1*], Luca Maria Aiello[2], Przemyslaw Grabowicz[3], Alejandro Jaimes[2] and Ricardo Baeza-Yates[2]

*Correspondence:
david.mabodo@gmail.com
[1]Universitat Pompeu Fabra, Barcelona, Spain
Full list of author information is available at the end of the article

Abstract

Social groups play a crucial role in online social media because they form the basis for user participation and engagement. Although widely studied in their static and evolutionary aspects, no much attention has been devoted to the exploration of the nature of groups. In fact, groups can originate from different aggregation processes that may be determined by several orthogonal factors. A key question in this scenario is whether it is possible to identify the different types of groups that emerge spontaneously in online social media and how they differ. We propose a general framework for the characterization of groups along the geographical, temporal, and socio-topical dimensions and we apply it on a very large dataset from Flickr. In particular, we define a new metric to account for geographic dispersion, we use a clustering approach on activity traces to extract classes of different temporal footprints, and we transpose the "common identity and common bond" theory into metrics to identify the skew of a group towards sociality or topicality. We directly validate the predictions of the sociological theory showing that the metrics are able to forecast with high accuracy the group type when compared to a human-generated ground truth. Last, we frame our contribution into a wider context by putting in relation different types of groups with communities detected algorithmically on the social graph and by showing the effect that the group type might have on processes of information diffusion. Results support the intuition that a more nuanced description of groups could improve not only the understanding of the activity of the user base but also the interpretation of other phenomena occurring on social graphs.

Keywords: social media; groups; bond theory; identity theory; Flickr

1 Introduction

The explosive success of social media is partly motivated by their capability of transposing everyday life dynamics on online platforms in a very intuitive way. Accordingly, even though dyadic social links are the primary way for people to connect online, social media have allowed from their very early stages the creation of social *groups*. This is a necessity that emerges directly from the collective behaviour of the crowd, that tends to flock in communities pushed by a number of reasons, including affiliation by similarity, local proximity, common interest, conflict with other groups, or even just the need for a definition of an identity by being separated by the rest of the population [1–4]. As a result, groups in

social media have flourished and they nowadays form a strong basis for user participation and engagement in online services.

For this reasons, online groups have been studied extensively in the past, with respect to their social structure and activity evolution. Despite the great attention given to the study of online groups, previous work on large online datasets has mainly considered groups as homogeneous entities, overlooking the fact that groups, similarly to social ties [5, 6], are not all created equal, as they emerge from different collective processes and from the different motivations of their founders or members.

Although several other disciplines, including physics, psychology, organizational sciences, and social sciences have been trying to explore specific aspects of the formation, evolution, and internal dynamics of groups at different levels (see Section 2), most of the studies has focused either on (i) small or offline social ecosystems, (ii) groups that are generated ad-hoc to conduct specific experiments, or (iii) groups inferred from the network structure. Also, very often specific aspects of group dynamics (e.g., consensus reaching, language norms, geographic placement) have been investigated in separation, with very few efforts to go towards a more holistic, multidimensional characterization of social aggregations. As a consequence, we feel that a thorough and large-scale exploration of the *nature* of online, user-generated groups, across some fundamental dimensions that characterize groups is in order.

We propose a categorization of online groups along three axes: *spatial*, *temporal* and *socio-topical*. For each dimension we propose a set of general metrics that capture quantitatively the different facets of groups. Specifically, we describe groups with respect of the geographic scattering of their members, the temporal footprint of the members' activity in terms of dispersion, skeweness, and burstiness, and the tendency of the group to aggregate members on a topical or social basis. With respect to the last dimension, we rely on a longstanding theory about the creation of social communities. The theory states that people join groups driven by either pre-existing social ties with other members or by the interest in the topical focus of the group as a whole and we build metrics to quantify this tendency. We show that our metric well reflect the cardinal points of the theory, being good predictors of the group type. Our metrics are tested on a large-scale corpus of public, online, user-generated groups.

Last, to frame our contribution in to a wider context, we provide examples of possible applications of our framework to other analytical issues on social networks. In particular, we put in relation the social and topical groups we find algorithmically with the communities detected from the graph structure and we speculate about the impact that different group types may have in the process of information diffusion, following the intuition that information cascades and group boundaries are strictly related concepts [7].

1.1 The Flickr case-study

We test our group characterization framework on a large scale set of online groups from Flickr (www.flickr.com). Flickr is a popular photo-sharing platform in which users can upload a large amount (up to 1 TB) of photos, organize them in albums or with free-form textual *tags*. Flickr provides means of rich social interactions between users. First, photos are shown in the user profile page and other users can *view* them, *comment* on them or mark them as *favorites*. Also, users can establish explicit social ties by *following* people they are interested in, to receive their status updates. Last, a pivotal part of the community engagement in Flickr is represented by *groups*.

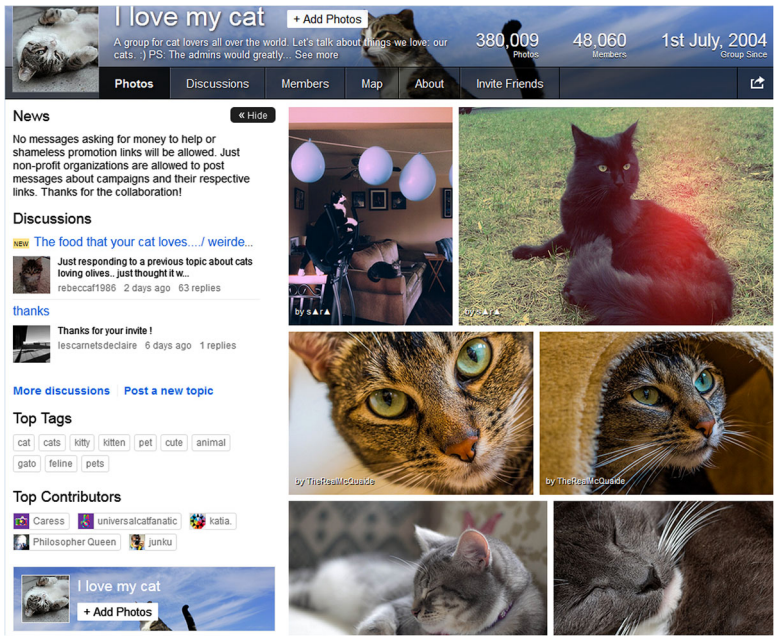


Figure 1 Page of a popular Flickr group, showing the pictures in the photo pool and a set of content (news, add photo, top tags) and social (discussions, members, top contributors) features.

Groups in Flickr are self-declared and self-managed communities that are spontaneously created by the user base, meaning that each user can create (and become administrator of) an arbitrary number of groups. Most groups have open membership, so users can join without invitation, just by clicking a *join* button, while others are by invitation only and joining requires the administrator's permission. Groups are usually built around a theme that is user-defined and, consequently, their topic, generality, and scope of interest can vary much. All groups provide functionalities that are explicit instantiations of *content* and *relational* features of social media. As illustrated in Figure 1, group participants can share their pictures with other group members in a common *photo pool*. A picture that is featured in a group tends naturally to receive feedback from other members, but social interaction is also possible through discussion boards. As a result, a Flickr group can be denoted by a set of *terms* including comments and tags on photos of the pool, discussions, group name and description and by a set of social *actions* such as exchange of messages in a discussion board, comments and favorites on photos, and so on.

Flickr groups represent an ideal ecosystem for the study of group characterization for a number of reasons. Groups in Flickr are large scale (hundreds of thousands of public groups, with a broad range in membership size), spontaneously generated (in contrast with groups inferred by the structure of the social network or created ad hoc for specific experiments), and exhibit public online information that is rich both in terms of content (photos, tags) and social information (multiple types of interactions between members). This combination of features is ideal to investigate the factors that drive the collective interaction between people in social aggregations. It is very difficult to find other large-scale, publicly accessible datasets with a similarly wide and diverse set of features. For these reasons, we focus our study on Flickr only, diving deep in several aspects of the groups' structure and organization rather than proposing a wider multi-dataset exploration.

1.2 Contributions and roadmap

This work is a direct extension to our previous paper [8] that focused on the interplay between social and topical aspects of online communities. Here we extend and improve that work here in a number of ways, and present the following contributions.

- We introduce a framework for the characterization of groups along geographical and temporal dimensions.
- We run a study of a large scale corpus of Flickr groups using the three target dimensions, being able to draw a more nuanced characterization of them than previous work.
- We use our framework to run a faceted analysis of the phenomenon of information diffusion on networks, spotting insightful correlations between type of spreading and type of group.

Overall, our work gives a contribution in the first place in the field of computational social science, specifically in the direction of a nuanced characterization of groups according to notions of topicality and sociality developed in sociology in the past decades but never tested on large online datasets. Our experimental evaluation shows that the formulation of the theory well captures the separation between the two macro-classes of groups. The transposition of the theory in quantitative metrics allowed us also to provide additional evidence to support another well-established theory about the maximum number of stable relationships for individuals in social environments (Dunbar's number [9]). Furthermore, we consider spatial, temporal, and socio-topical metrics jointly for the first time, discovering some macro-classes of groups that reveal the interplay between the different dimensions; to mention two clear examples, topical groups that tend to be long-lived and with steady activity in contrast with social groups that are more often bursty and short-lived.

The main goal of this work is to provide yet another step towards a computational understanding of social structures, user-generated groups in this specific case. To show that the value a nuanced characterization of social aggregations is not limited to the possibility of carrying out more fine-grained network analysis, we also connect our study to the field of information diffusion showing that different types of groups can impact on the process of the spreading of information along the network. This is the first study that shows such empirical evidence and directly connects with very recent work in information diffusion that have been trying to leverage the same intuition [7].

The remainder of the paper is structured as follows. First, we present an overview of related work (Section 2). Then we introduce the three dimensions that we use to characterize social groups (geographical, temporal, and socio-topical) and we define how to measure them quantitatively (Section 3). After a short illustration of the Flickr dataset we use and of the ground truth we extracted to validate our socio-topical metrics (Section 4), we present the results of the application of the metrics, identifying different classes of Flickr groups with respect to the three dimensions considered in separation but also jointly (Section 5). Finally, we set our contribution into a wider context by analyzing the process of information diffusion in the light of the different group types in which the process takes place (Section 6).

2 Related work

2.1 Online groups characterization

Since the very early stages of the social web, the research community has been interested in the definition of the notion of group and of its possible types [10] not only for analytical

purposes but also in direct application to several tasks, including profiling and recommendation [11, 12]. The global structure, evolution and dynamics of social groups have been investigated over large-scale and heterogeneous datasets. The shape and evolution of groups have been described in computer science literature as very broad phenomena [1, 13] that are determined by the intrinsic group fitness [14] and on the density of social links connecting their members [15].

Although the broad variety of group types and their emerging features (starting from their size [16]) has motivated some research work to characterize the nature of groups along their main dimensions, most of the contributions so far have not established any quantitative framework for their classification.

Due to its open nature and its multiple features, Flickr has been one of the most studied platform to this respect. Early work relied on interviews and user studies to identify the different usage of Flickr groups [17], finding five main motivations for users to join groups (memory, identity and narrative, relationships maintenance, self-representation and self-expression). Alternative classifications based on user studies have been proposed as well [18, 19].

Negoescu *et al.* have contributed quite much to this research area with several studies on Flickr groups. First they have introduced a manual categorization of Flickr groups, partitioning them in *geographical*, *topical*, *visual*, and *"catch-all"* groups [2]. With this categorization in mind, they propose to detect hypergroups (i.e., groups of groups) based on the similarity of their topical focus, extracted with LDA [20]; in contrast, Negi *et al.* try to find subgroups in large Flickr communities using MoM-LDA on photo tags [21]. Groups have been also studied in relation with their membership, with special attention to topicality and to recommendations exchanged between peers [22]. In more recent work [23] Negoescu *et al.* have discussed about how to represent Flickr groups group according to the topics and tags in use by their members. Also, according to previous studies [17], they identified "real" groups as those motivated by self-expression and relationship maintenance. However, although every Flickr group can be mapped to a topic (set of words), not all groups have a topical focus, as we show in this work.

Following an earlier conceptual framework [24], Cox *et al.* [13] attempted to measure the "groupness" of a group using several metrics as size of membership, volume of photos, length of description, and so on. They propose a classification of groups into topical (focused on a theme), highlighting (to promote photos to a wider public) and geographical (rooted into a specific geolocation); however their classification is ultimately arbitrary and not supported by quantitative results. In partial contrast with previous work [23], their results also point out that small groups are more important than the big ones to the social activity of the network as they operate at "human scale". The work was subsequently extended [25] and the categorization was manually refined into four categories, namely location-based, award, learning, an topical groups.

Prieur *et al.* use Pricipal Components Analysis (a statistical procedure that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables) on a set of features extracted from Flickr groups to detect the main dimensions that characterize them [26–28]. They find three main dimensions underlying as many types of groups: social media-use, MySpace-like, and photo stockpiling. The mixture of sociality and topicality of groups is also discussed, even though only tangentially.

At a finer scale, social communities can be described in terms of user engagement. From a quantitative perspective, the amount of participation of members in activities related to the group is varied and dependent on group size [29]. Intra-group activity has been characterized in terms of propensity of people to reply to questions of other members [30], coherence of discussion topics [31], or item sharing practices [2]. Modeling inner activity of groups has helped in finding effective strategies to predict future group growth or activity [3], recommend group affiliation, or enhance the search experience on social platforms [20].

Groups have been studied also in other online platforms. The structure of user interaction patterns in groups extracted from LiveJournal, DBLP, YouTube, Orkut, and Yahoo Groups have been investigated in the past [1, 29, 32, 33]. Laine *et al.* [34] present an analysis on YouTube groups, highlighting their tendency to both topicality and sociality and the small-world nature of the interactions inside them. Interestingly, they envision in future work an analysis of the interplay between groups and influence.

Last, some important contributions in the field of complex systems have investigated some properties of groups that can be inferred by the structure of social links via clustering or community detection algorithms. Barabasi *et al.* [35] use a network of phone calls to find that large detected groups (using clique percolation) persist longer when they are capable of dynamically changing their membership, suggesting that the ability to self-altering the internal composition results in better adaptability. Using a similar dataset, Onnela *et al.* [36] explore the geography of groups to find that small communities are geographically tight, but become geographically wider when the group size exceeds 30 members. Temporal patterns similar to the ones we explore in this work have been also investigated. Burstiness of human behaviour as a consequence of processing tasks in relation with their perceived priority have been studied, but not specifically in the context of groups [37].

2.2 Groups in (computational) social sciences

Small-size groups have been studied by researchers in psychology and sociology and by scholars in other social and behavioral sciences for the past century and especially in the past decades. The notions of *community* and social *group* have been widely debated in behavioral sciences [38, 39]. The faceted complexity of groups, have been discussed for long time [40] and previous works have remarked that the internal dynamics of social groups emerge from the combination of complex cognitive processes such as sense of membership, influence between people, fulfillment of individual and collective needs inside the group, and shared emotional connections [41]. Based on such widely accepted theoretical foundations, sociological theories have been formulated to disentangle all of these complex aspects.

From the perspective of organizational sciences, groups have been investigated with special focus on computer mediated communication, namely how belonging to a group can affect the communications between members, in time. Siegel *et al.*, for instance, have run a small-scale comparison between online and offline groups, showing that the process of finding a consensus is resolved with a significantly bigger shift from the initial members' opinion in the online case rather than the offline one [42]. Also, other studies highlighted that members physical co-presence is an important factor for the success of a task-oriented group, as the geographical distribution of members could negatively impact the effectiveness of communication [43]. Similar patterns are found also when considering the time

dimension, as "ongoing" teams that have to collaborate for a long time must tackle more process and structural issues than groups with "temporary" tasks [44]. In environments where many groups co-exist, overlap between with other groups' memberships, age and size of the group can be factors that affects the group in terms of its growth, as result of competitive pressure [45]. Also, individual benefit of group members in collaborative tasks can be greatly driven by the impulse given by the group leader to increase community building [46]. The use of language within social communities has been studied by Postmes *et al.*, who look at how interactional norms regulate the conversation style (e.g., use of abbreviations, superlatives) in small email discussion groups, shedding light on the processes of social construction and formation of social identity [47].

More recent work in computational social science attempted to characterize groups in relation to well-established theories from social sciences. The dependency of activity and connectivity on group size has been studied in several platforms [3, 48, 49], showing relations to Dunbar's theory on the upper bound of around 150 stable social relationships for an average human [9]. The dimension of similarity between members has been identified also as a factor driving the creation of social communities [50], particularly given that, to a large extent, users in social networks tend to aggregate following the homophily principle [4]. However, similarity is not necessarily the strongest indicator for group activity and longevity, as diversity of content shared between group members is a major factor to keep alive the interest of members [51].

Social and thematic components of communities have been widely studied in social science, most of all within the common identity and common bond theory on which part of the present work is based [52–54]. Nevertheless, the principles behind the theory have never been translated into practical methods to categorize groups, nor tested on large datasets. On the other hand, data-driven studies have investigated social and thematic components separately when characterizing groups [13]. Preliminary insights on the interweavement between such dimensions have been given in exploratory work on Flickr, where signals of correlation between social density and tag dispersion in groups is shown [26] and where two different clusters emerge naturally when plotting the groups size against the number of internal links [16]. In this work, we define metrics that can be used to predict if a group is social or topical and testing their effectiveness against a reliable ground truth.

2.3 Automatic group extraction

Besides the analysis of user-created groups, the study of automatically detected groups through community detection algorithms has attracted much interest lately [55]. Detected communities are supposed to represent meaningful aggregations of people where dense or intense social exchanges take place among members [48]. Nevertheless, even if synthetic methods to verify the quality of clusters have been proposed [56], the question of whether such artificial groups capture some notion of community perceived by the users remains open. If on the one hand the computation of cluster-goodness metrics over user-created groups can give useful hints about their structural cohesion [57], on the other hand a direct comparison between user-created groups and detected communities is still missing, particularly in terms of the amount of sociality or topical coherence they embed.

2.4 Information propagation

Modeling the dynamics of information diffusion and influence along network links has received much attention in the last decade, especially in relation to the task of optimization of viral marketing strategies [58]. A large corpus of studies on influence and information propagation has relied on Twitter-based experiments [59, 60]. In Flickr, instead, an analysis of information propagation based on favorites showed that diffusion is limited to individuals who reside in the close neighborhood of the seed user and the spreading process is very slow [61]. In this paper we use the same propagation model to measure the effect that the boundaries of different groups may have on diffusion of information. Instead of representing the influence as an infection phenomenon between connected individuals, alternative models agnostic on the network structure that rely only on the time of the contagion have been proposed [62, 63], assuming the presence of a hidden contagion web that might be different from the observed social network [64].

Even though our contribution does not focus on the definition of information propagation models, we measure information diffusion through social links within different group types, motivated by recent findings that hypothesize a connection between group type and potential of propagation of information cascades [7, 65].

3 Metrics for group characterization

Geography, time and the duality between social and topical bias of groups are mentioned multiple times in previous work, as they are important aspects for the characterization of communities. Next, we consider those three dimensions and define new general metrics for each of them. All our metrics assume the presence of a user base U and a set of groups G where $g = \{u_1, \ldots, u_n\}, g \in G \wedge u_i \in U$. Users can belong to multiple groups and we associate, with each group, a bag of user-generated *terms* T_g (e.g., tags, group posts). We also assume to have a set of actions A_g that members of a group g perform within the group (e.g., group subscription or photo upload in the group pool). In the following, we refer to these actions also as events. We consider space and time associated to those actions, respectively geo = $(\text{lat}_a, \text{lon}_a)$ and $t(a) \equiv t_a$, $a \in A_g$; when focusing on a specific type of action, we will consider their temporal sequence, whose timestamps will be denoted simply as t_i, $i \in [0, n]$. Last, we take into account also the social interactions between members of a group and within groups. We adopt a very general *multidigraph* model that fits most of the current social media platforms. Members are represented as nodes, and each distinct *interaction* between any two members is represented by a directed arc.

3.1 Geographical dispersion

Geographical distribution of group members can be very variable, as sometimes groups are very localized (e.g., members of a photography club in the same city) and sometimes very broad (e.g., Canon camera owners). When studying the geographical distribution of viewers of a given photo, Van Zwol [66] proposed three metrics. First, the most direct way to gauge the sparsity is to compute the geodesic distance geo_d between all the pairs of locations

$$\begin{aligned}&\text{geo}_d(\text{lat}_1, \text{lon}_2, \text{lat}_2, \text{lon}_2)\\ &= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\text{lat}_1 - \text{lat}_2}{2}\right) + \cos(\text{lat}_1) \cdot \cos(\text{lat}_2) \cdot \sin^2\left(\frac{\text{lon}_1, \text{lon}_2}{2}\right)}\right),\end{aligned}$$

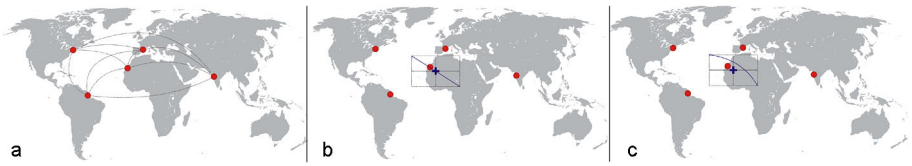
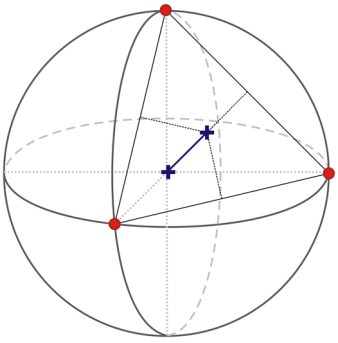


Figure 2 Illustration of methods to measure dispersion of geolocated points (in red) on a map. **(a)** Average of pairwise geodesic distances. **(b)** Diagonal of the bounding box defined by the standard deviations of latitude and longitude around the center of gravity (blue cross). **(c)** Same as **(b)** but with geodesic distance of the diagonal.

Figure 3 Illustration of the center-to-Earth distance method to measure dispersion of geolocated points (in red) on the Globe. Points are translated into spatial Cartesian coordinates, they are averaged, and the distance of the resulting point to the center of the Earth (in blue) is calculated as a measure of dispersion.



and to average them (Figure 2(a)). However, this scales quadratically with respect to the number of points and it could be computationally prohibitive when large sets of points are considered. For this reason, a second way to estimate dispersion is to compute the standard deviation for the longitudes and latitude separately and use them to build a bounding box around the centroid of the Cartesian coordinates (Figure 2(b)). Then the Euclidean distance between the angles of the bounding box is considered ad a measure of spreading. This solution however does not consider the rounded surface of the Earth, thus biasing the results by the latitude: same values at different latitudes could imply very different distances. A straightforward solution to solve this problem is to user the geodesic distance instead of the Euclidean distance (Figure 2(c)). Still, even if this solution accounts for curvature, it does not consider the Earth as spherical, as longitude is interpreted as a linear metric (e.g., two points at the two ends of the Bering strait will be considered very far from each other).

To address these problems we propose a new simple metric, the center-of-Earth distance (coe_d), to directly measure geographical dispersity, illustrated in Figure 3. We consider each latitude-longitude pair as a polar-azimuth angle in the spherical coordinate system centered on the center of Earth. We convert all the points into the Cartesian system, so that every point is represented as points in the three-dimensional space. All the points are then averaged and, as they all lie on the spherical surface, their average will be a centroid that by definition will be under the Earth's surface. The sparsity is then estimated by the distance of the centroid to the center of the Earth, normalized by the Earth's radius. In the case-limit in which just one point is available (or many perfectly overlapping points), the geographical spreading will be maximally narrow ($\text{coe}_d = 1$), whereas two points at the antipodes will have a centroid residing exactly at the center of the Earth ($\text{coe}_d = 0$), thus

leading to maximal sparsity. Additionally, we apply the arc-cosine to the final value to get an angle that more intuitively relates to the spreading of points on the spherical surface.

This solution is linear with the number of points, it considers the Earth's curvature and it considers the World as spherical, thus addressing the limitations of previous approaches.

3.2 Temporal patterns

Similarly to geography, groups could exhibit also quite broad temporal patterns. The time series of events associated to a group (e.g., photo uploads) is the temporal footprint we aim to characterize. Of course, as each distribution in time is likely to be unique, we need to capture the peculiar features of each temporal pattern. We rely on the statistical properties of the distribution of the volume of actions in time to describe the time sequences. We identify four different properties: the *central tendency*, the *dispersion*, the *skewness* and the *burstiness*. In the following, we consider that all the events take place in a fixed, large time window $[0, T]$ (that will correspond to our temporal sample in the experimental data). Next, we define their meaning and propose metrics to capture each of them. The way to combine the metrics for a characterization of groups along the temporal axis will be discussed in Section 5.2.

3.2.1 Central tendency

We consider a sequence of timestamps $(t_0, \ldots, t_n)$ in which events in the group occur. For each group, we consider a normalized window where the start time is 0 and coincides with the first group event ($t_0 = 0$) and ends at time 1, which coincides with the end T of our temporal sampling (which implies $t_n \leq 1$). We define the central tendency of a group as:

$$\mu_g^t = \frac{1}{n} \sum_{i=0}^{n} t_i. \tag{1}$$

The output value is in the range $[0,1]$ and reflects the central tendency of distribution of events in time: the closer the value to 0 the most time values happened at the beginning of the group's life, the closer to 1, the most values near to the present. Groups with strong central tendency will have values close to 0.5.

3.2.2 Dispersion

Dispersion denotes how stretched or narrow is a distribution. To measure dispersion we use a corrected version of the standard deviation using events in a normalized timeline:

$$\sigma_g = \sqrt{\frac{1}{n-1} \sum_{i=0}^{n} \left(t_i - \mu_g^t\right)^2 \frac{1}{n(1-\mu_g^t)\mu_g^t}}. \tag{2}$$

Values range from 0 to 1. Note that groups with high central tendency would have low dispersion, but groups with low dispersion could have also low central tendency. However, a non-corrected standard deviation would still be dependent by the central tendency, as for instance a series of time events with central tendency value of 0.1 cannot have a dispersion higher than 0.5. To ensure that the independence between metrics the correction value is required. For the sake of brevity, we do not report the mathematical details here, but a mathematical justification of the correction is reported in the Appendix.

3.2.3 Skewness

Another of the time properties we want to capture is skewness. It is calculated with the normalized difference between the median and the mean as follows:

$$\gamma_g^t = \frac{\mu_g^t - \text{median}_g^t}{\min(\mu_g^t, 1 - \mu_g^t)}. \tag{3}$$

Again, values are in the $[0,1]$ interval. A divergence between the mean and the median implies a skewed distribution as more elements will have values either smaller or larger than the median. The correction factor in the denominator ensures the independence between the skewness and the central tendency, as shown in the Appendix.

3.2.4 Burstiness

Last, we use a measure of burstiness to check whether the events are grouped together and happen in big bursts. To capture this concept we consider the inter-event time $\Delta_g^{t_{ij}} = t_j - t_i$, $i < j$ is used. We denote as Δ_g^t the overall series of inter-event times for group g. The burstiness is defined as follows:

$$\Delta = \log_{10}\left(\mu\left(\Delta_g^t\right)\right) - \log_{10}\left(\text{median}\left(\Delta_g^t\right)\right). \tag{4}$$

Note that the mean of all the inter-event times $\mu_{\Delta_g^t}$ is equivalent to the total time between the t_0 and t_n, divided by the number of events. The median of the inter-event times instead will get values on the range $(0, \mu_{\Delta_g^t})]$. For the series with uniformly separated events, $\mu(\Delta_g^t)$ and $\text{median}(\Delta_g^t)$ will be equal whereas the groups with a bursty behavior will have a $\text{median}(\Delta_g^t)$ near to 0.

3.3 Topical and social groups

The "common identity and common bond" theory [52] states that, depending on the prevalent motivation of people to join a group, groups can be categorized as either *social* or *topical*, and assumes that the two types of groups have distinct and well-recognizable traits. In recent years, the theory has been widely commented and elaborated by social scientists from a theoretical perspective and through small-scale experiments [53, 54, 67], but no rigorous methodology to distinguish the two types has been developed nor tested on large-scale datasets.

We design a technique to detect the group type based on the common identity and common bond theory, first to contribute to a strong validation of the theory itself but also to provide a general framework for automatic classification of user groups in online social media. In the following, we provide a more detailed description of the theory and then we propose a translation of its main principles into general metrics that can be applied to social graphs.

3.3.1 Identity and bond theory

The *common identity and common bond theory* describes social groups along the dimensions of topicality and sociality [52, 54]. According to the theory, the attachment to a group, as well as the permanence and involvement in it, can be explained in terms of common *identity* or common *bond*. Identity-based attachment holds when people join a group

based on their interest in the community as a whole or in a well-defined common theme shared by all of the members. People whose participation is due to identity-based attachment may not directly engage with anyone and might even participate anonymously. Conversely, bond-based attachment is driven by personal social relations with other specific members, and thus the main theme of the group may be disregarded. The two processes result in two different group types, that for simplicity we name "*topical*" for identity-based attachment and "*social*" for bond-based attachment.

In practice, groups can be formed from a mix of identity and bond-based attachment, but very often they tend to lean more towards either sociality or topicality. According to the theory, the group type is related with the *reciprocity* and the *topics* of discussion. Members of social groups tend to have reciprocal interactions with other members, whereas interactions in topical groups are generally not directly reciprocated. In addition, topics of discussion tend to vary drastically and cover multiple subjects in social groups, while in topical groups discussions tend to be related to the group theme and cover specific areas. According to the theory, social groups are founded on individual relationships between their members, therefore it is harder for newcomers to join and integrate with members that already have strong relationships between each other. One implication of this is that social groups are vulnerable to turnover, since the departure of a person's friends may influence his own departure. Topical groups, on the other hand, are more open to newcomers and more robust to departures.

3.3.2 From theory to metrics

It is possible to construct metrics to differentiate between the two types of groups by quantifying the reciprocity of interactions, and the topicality of the information exchanged between group members. Next, we describe: (i) *reciprocity* metrics, used to quantifying group sociality, (ii) *entropy* of terms, to determine how much the topics of discussion vary within a group, and (iii) *activity* metrics, to measure the liveliness of the group. Similarly to the temporal dimension, the approach to combine all these metrics into a decision on the group type will be discussed in Section 5.3.3, with specific examples on our Flickr case-study.

Reciprocity Reciprocity occurs whenever a user interacts with another user and that user responds her at any time later with the same type of interaction. We define *intra-reciprocity* of a group g as:

$$r_g^{\text{int}} = \frac{E_g^{\text{int,rec}}/2}{E_g^{\text{int,rec}}/2 + E_g^{\text{int,nrec}}}, \tag{5}$$

where $E_g^{\text{int,rec}}$ and $E_g^{\text{int,nrec}}$ are, respectively, the number of reciprocated and non-reciprocated links internal to the group g. Correspondingly, the *inter-reciprocity* at the border of the group is defined by r_g^{ext}, accounting for the reciprocity between members and non-members.

We normalize the intra-reciprocity score using the average reciprocity value $\langle r_g^{\text{int}} \rangle$ over all groups:

$$t_g = \frac{r_g^{\text{int}}}{\langle r_g^{\text{int}} \rangle}. \tag{6}$$

The larger the intra-reciprocity, the higher the probability that the group is social. Alternatively, to compensate for the effect of the correlation between reciprocity and the number of internal interactions, and to account for local effects, the intra-reciprocity can be normalized by the inter-reciprocity:

$$u_g = \frac{r_g^{\text{int}} + 1}{r_g^{\text{ext}} + 1}. \tag{7}$$

We add 1 to both numerator and denominator to reduce the fluctuations of u_g at low values of r_g^{ext}. This relative reciprocity compares the reciprocity between the members with their reciprocity toward people not belonging to the group. It reflects how sociality of group members distinguishes itself from the environment.

Topicality The set of terms $T(g)$ associated with a group indicates the topical diversity of the group. Thus we measure the entropy of the group as

$$H(g) = -\sum_{t \in T(g)} p(t) \cdot \log_2 p(t), \tag{8}$$

where $p(t)$ is the probability of occurrence of the term t in the set $T(g)$. The higher the entropy, the greater is the variety of terms and, according to the theory, the more social the group is. Conversely, the lower the entropy, the more topical the group is. In addition, since not all groups have the same number of terms and the entropy value grows with the total number of terms, we introduce the *normalized entropy* h_g, which is normalized by the average value of entropy for the groups with the same number of terms:

$$h_g = \frac{H(g)}{\langle H(f) \rangle_{|T(g)|=|T(f)|}}. \tag{9}$$

Activity Even if, for the considered theory, activity is not a discriminative factor between social and topical groups, it is useful to characterize the liveliness of a community. Activity is quantified in terms of the number of internal interactions normalized by the expected number of internal interactions for a set of nodes with the same degree sequence:

$$a_g = \frac{E_g^{\text{int}}}{(D_g^{\text{in}} D_g^{\text{out}})/E}, \tag{10}$$

where D_g^{in} and D_g^{out} are total numbers of interactions originated by members of the group g or being targeted to members of this group, where E is the total number of interactions in the network. If this property has a value higher than 1 then the number of interactions internal to the group is higher than the number of interactions expected in a random scenario with the same group activity volume.

Another way of measuring activity of a group is by comparing density of its internal interactions with the density of its external interactions:

$$b_g = \frac{E_g^{\text{int}}/(s_g(s_g-1))}{E_g^{\text{ext}}/(2(N-s_g)s_g)}, \tag{11}$$

where s_g is the cardinality of group g and N is total number of nodes in the network. Values of b_g greater than 1 indicate a density of internal interactions higher than interactions between the group and the rest of the network. This metric effectively compares intensity of interactions between members of the groups with the intensity of their interactions with the entire network.

4 Dataset and preprocessing

To test our metrics we use a dataset from Flickr. The wide variety of user groups, the richness of interaction types, and the openness of the data (retrievable through the public API) make Flickr an ideal platform for our study.

4.1 Flickr groups and interactions

We consider a random sample of public groups created until the end of year 2008. Users of Flickr can create, moderate and administer their own groups. Most groups are open, so users can join without an invitation. Others are only by invitation and joining requires the administrator's permission. In total, our dataset contains over 500K groups. For each of these groups, we extracted all the public information related to them (retrievable via the Flickr public API) and that we detail next. All the data has been anonymized and processed in aggregate. Table 1 summarizes some statistics of the data described below.

First, for all the users of the groups, we collect public information of their profile, extracting their interactions with other users or online objects, namely:

- *Comments*. User u comments on a photo of user v. This interaction is *mediated* through the photo. We filter out the comments of users on their own photos, obtaining a total of 238M comments.
- *Favorites*. User u marks one of user v's photos as a *favorite*. The interaction is mediated through the favorited photo. We extract 112M favorite interactions.
- *Contacts*. User u adds user v among his contacts. Social contacts in Flickr are directed and may be reciprocated. One person can choose another person as his contact only once and the relation remains in the same state until the contact is removed. There are 71M contacts in our dataset.

Additionally, we also rely on the information related to specific actions that users make to interact with the group itself:

- *Uploads*. User u uploads a photo p to the group *photo pool*. Flickr groups provide pools to store pictures related to the group and pictures can stay in multiple pools. Only members of the group can upload a photo to a pool.
- *Subscriptions*. User u joins a the group at a certain time.

In addition to user-created groups (we refer to them as *declared*), in Section 5.1 we analyze the sociality and topicality properties of groups that are not defined by users but are instead found by community detection algorithms (we name these *detected* groups). We applied the OSLOM community detection algorithm [68] over the entire network of social contacts in our dataset. We choose OSLOM because it detects overlapping communities, which is a natural feature of real groups. Moreover, OSLOM has performed well in recent

Table 1 Total number of interactions and declared/detected groups.

Comments	Favorites	Contacts	Decl. g.	Det. g.
238M	112M	71M	504K	646K

community detection benchmarks [56] and it outperformed other algorithms we tested. OSLOM detected 646K groups.

We also use *tags* of the photos as terms for our model. The primary set of photos from which we extract tags is the photo pool. Photo pools are available for declared groups only. In addition, in both declared and detected groups, the interactions between members of the group that are mediated through photos (i.e., comments, favorites) result in two additional photo sets from which tags are extracted. We process the three tag sets separately (pool, comments, favorites), and for each of them we compute the normalized entropy (h_g^{pool}, h_g^{com}, h_g^{fav}).

4.2 Socio-topical group labeling

The socio-topical dimension we consider is a rather abstract concept and we like to check whether our metrics are able to correctly capture it. For this reason, we need a reliable ground truth to check against the detected sociality and topicality scores. We asked human coders to label groups based on well-defined guidelines extracted directly from the common identity and common bond theory [54]. For the labeling we randomly selected groups meeting the following minimum requirements: (i) more than 5 members, (ii) more than 100 internal comments, (iii) relative activities a_g^{com} and b_g^{com} higher than 10^2. The third requirement ensured us that the selected groups were active well above the expected values in a random case. After this selection we obtained over 34K declared groups and over 33K detected groups. We describe the labeling process of such groups in detail next.

4.2.1 Information provided to labelers

The labeling is based on the human capability of processing the semantics, aesthetics, and sentiment behind text and photos. With the editorial process we generate a ground truth of "social" and "topical" groups. The coders were asked to make judgments in this respect and were presented with the following information for each group:

Group profile. The Flickr group profile consists of the group name, description by the creator of the group, discussion board, photo pool, and map of places where photos uploaded to the group pool were taken. This information is available only for declared groups.

Comments. We provide text of all comments that happen between the members. Comments are shown in chronological order and are grouped by thread, if they appear under the same photo. Additionally we also include a link to the photo.

Tags. Human coders are shown the list of the 5 most frequent tags attached to the photos that mediate the internal comments to the group. The list is sorted alphabetically.

4.2.2 Labeling guidelines

Coders were shown the information described above and asked to categorize groups as either *social*, *topical* or *unknown*. The last case is reserved for groups for which text is written in a language unknown to the labeler, making the task impossible to accomplish. Intentionally, no *unsure* category was allowed to keep the categorization strictly binary, as the theory does. Some groups can be both topical and social, and therefore difficult to categorize, but for the sake of clarity and conformity with the theory we kept the categorization as a binary task. Coders were provided with specific instructions on how to recognize social and topical groups, and on how to perform the categorization. The guidelines are summarized as follows:

I. Comments and photos. By examining comments and photos, find traces of people who know each other or who have a personal relationship. Knowing each other's real names, spending time together, co-appearing in photos, sharing common past experiences, referencing mutually known places, and disclosing personal information are all signals of the presence of a social relationship [69]. The predominance of friendly and colloquial comments (e.g., jokes, laughter) is another element distinguishing social groups from topical groups. In topical groups, the atmosphere is more formal and comments tend to be more impersonal [53]. Examples of impersonal comments include expressing appreciation for photos, praising the photographers, thanking them for their work, or commenting on any particular topic in a neutral way. As a rule of thumb, if many personal comments are detected, then the sociality of the group should be considered high. If such comments are not many (e.g., just between small subsets of members), but the overall atmosphere of the interaction is rather personal and friendly, then we consider the sociality of this group as fairly present. If, on the other hand, comments are mainly impersonal and neutral, sociality has to be considered low, in favor of higher topicality.

II. Tags and description. Read the tags and the profile description of the group. If the tags are semantically consistent then the topicality of the group should be considered high, and even higher if the name and description of the group corresponds to the content of the tags. In some cases, tags or group descriptions can contain words indicating personal relations or events (e.g., "wedding", "grandpa", names, *etc.*), indicating a higher sociality of the group. Tags can also contain names of specific locations. Geo-characterized tags can be reasserted by looking at the map of places where photos were taken. Such tags are a good indication that the sociality of the group is present, but that has to be confirmed through the inspection of comments.

The coders labeled the groups after judging the two aspects above. If both tags and comments are highly social or topical, then the choice of label is straightforward. If the tags are highly topical and the comments are not social then the group is labeled as topical, and vice versa. If the tags are a bit topical and comments highly social then the group is labeled as social. The labelers were asked to read as many comments as needed to arrive to a fairly clear decision.

4.2.3 Group examples

To provide a sense of how the defined guidelines were applied in practice, we describe two examples. The first one is a group titled "Airlines Austrian", tagged with labels "aircraft", "airport" and "spotting." Photos are from different countries in Europe and the vast majority of them depict airplanes. Members are very active in commenting and writing comments related on the aircraft theme (e.g., "I just love this airplane, the TU-154M is just a plane Boeing or Airbus could never design"). In this case, all of the features are aligned with the concept of topical group defined in the guidelines. The second group is named "Camp Baby 2008" and it is described in the main page as a collection of photos of a two-day event for young mothers taking place at a specific location. Photos depict people attending the event and interacting with each other with a friendly attitude. Tags and comments often contain names of individuals and references to past common experiences (e.g., "I love Mindy and cannot wait to see her again!!"). Although the group has a specific topic, its social component is very strong. In practice, more ambiguous cases can occur and, ultimately, the decision of the labeler has an arbitrary component, as in

every complex annotation process. Nevertheless, the defined guidelines gave the labelers precise instructions and, as described next, we recurred to multiple independent coders to assess the quality of the extracted ground truth.

4.2.4 Labeling outcome

A total of 101 declared groups and 69 detected groups were labeled by 3 people: two of the authors and an independent labeler who was not aware of the type of study nor of the purpose of the labeling. The inter-labeler agreement, measured as Fleiss' Kappa, is 0.60 for the declared groups, meaning that there exists good agreement between labelers.

In order to assess the quality of the labels, we also counted the number of messages exchanged between group members. The counting was done anonymously in aggregate and the content of the messages was not accessed. Groups labeled as social contain around twice as many messages between their members compared to topical groups of similar size. Even if this does not constitute a proof of higher sociality, intuitively people who get in touch via one-to-one communication are more likely to have a more intimate social relationship.

The Kappa value for detected groups is around 0.44, revealing lower agreement. A factor that partially determined such result is the lack of information about the group's profile, since it is not available for detected groups. Another cause of the disagreement is a higher variability in the comments. This may be because we use a network of contacts for the purpose of finding clusters and defining detected groups, which may not be the best proxy of personal relations.

In total we label 565 distinct declared groups and 126 distinct detected groups. We characterize them in the following section.

5 Characterization of groups

We now describe the Flickr groups in our dataset according to the three dimensions identified above. After a short analysis of the overlap between declared and detected groups, we inspect each dimension separately, discussing how the metrics we identified earlier are applied to groups. Last, we discuss the characterization of groups along all the three dimensions.

5.1 Overlap of groups with detected communities

Since community detection techniques have been largely employed in recent years to describe the structure of complex social systems [55], the need for a clearer assessment of the meaning of the detected clusters has been often expressed from different angles [56, 57], but never completely satisfied. In this study we contribute to shed light on this matter by comparing the user-generated groups with the groups detected algorithmically (as described in Section 4).

The groups from the two sets share typical properties of groups found in on-line social networks. The distribution of sizes of groups in both cases is heavy-tailed and close to power-laws (not shown). Declared groups tend to be much bigger, having on average 61 members versus 7 members in detected groups.

To test if the groups from the two sets overlap, and to what extent, we measure the Jaccard similarity between their sets of members. Similarity is computed for all declared-detected group pairs and for each detected group we select the declared one with the highest similarity value as the best match. We plot the average similarity of the best matches as

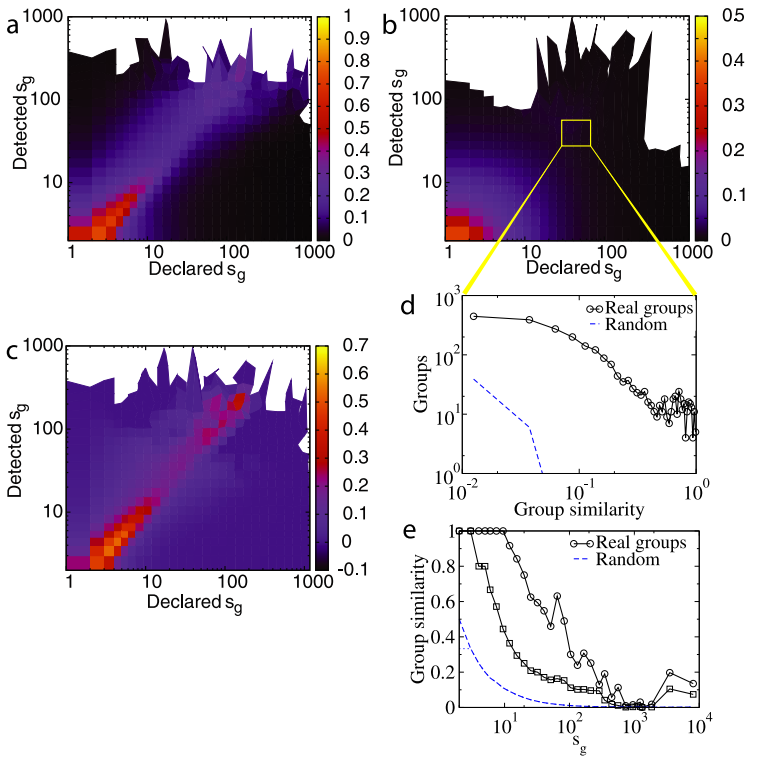


Figure 4 Jaccard similarity between declared and detected groups as a function of their sizes. Diagonal shows an interesting pattern **(a)** which is not reproduced by randomized groups **(b)**. We subtract **(b)** from **(a)** and plot the result in **(c)**, and a histogram of similarity values for a sample of groups lying at the diagonal in **(d)**. For groups of various sizes, we plot 91th and 99th percentiles of similarity between declared and detected groups **(e)**.

a function of the size of groups in Figure 4(a). Zero values of similarity are not taken into account for these averages. For the purpose of comparison with a null model, in Figure 4(b) we draw the same plot after randomly reshuffling the members of detected groups, while preserving their sizes. We observe that the two plots differ in values significantly along the diagonal, and that the difference between them is substantial, as shown in Figure 4(c), meaning that indeed detected groups are, to some extent, similar to the declared ones. Further insights are shown in Figure 4(d), where we depict the distribution of similarities of pairs of groups extracted from a small sector of the diagonal, having between 32 and 64 members. The figure shows that there exist multiple detected groups which overlap significantly with declared groups, and that randomized groups do not show this pattern.

This holds for groups of all sizes, as shown in Figure 4(e), in which we plot the 91th and 99th percentiles of the best match similarity for detected groups of various sizes (e.g., 1% of detected groups of size 20 have similarity with declared groups higher than 0.75, while for the randomized case 1% of the groups have similarity higher than just 0.05). Therefore, in some cases the community detection algorithm finds groups which are also defined by users (i.e., declared groups). We present evidences that this does not occur by chance through the comparison with the randomized case. Nevertheless, a substantial overlap is found for just a small percentage of groups. Most of the group pairs have similarity close to 0. Consequently, the similarity of detected groups to the best-matching declared groups is 0.082, while for the randomized detected groups it is not much lower, yielding 0.058.

5.2 Spatio-temporal classes

Spatial characterization of groups is defined by a single dispersion metric coe_d. In Flickr groups we have two potential different sources of geolocated data: user location and photo geotags.

Here we do not use the geolocations of users for two reasons. First, some users do not provide their position and the IP-based geolocation could be quite unreliable [66]. Last, we aim to characterize groups with the information that is directly related to that group rather than to an individual. For this reason, we consider the geotags attached to the photos uploaded to the group instead.

When we apply the dispersion metric using geotags, we obtain a distribution over groups that is shown in Figure 5. The histogram displays a bi-modal distribution: a first local maximum around zero that includes the groups that contain photos geographically near, and a second local maximum with peak around the 0.85 radians ($\approx 50°$), that is approximately the angle between Europe and US, which are the two continents with highest data density. A random sample of photos in the dataset produces a peak at the same point (not shown), therefore suggesting that groups with those higher dispersion values are groups where the geographical aspect is not crucial to the purpose of the community.

To transition from a continuous value to a partition of groups into classes we apply the X-Means algorithm [70] over the monodimensional space of dispersion values, to avoid manual thresholding. X-Means is an improvement over K-Means where the number of clusters K is not given and it is able to estimate the number of clusters and the clusters in a much faster way than optimizing the parameter K with brute force approaches.

Not surprisingly, two clusters are found. The *geo-narrow* cluster, contains the 56% of groups, and the remaining 44% belongs to the *geo-wide* cluster.

The temporal aspect includes four different metrics that would be difficult to combine with ad-hoc approaches. Besides, we have two different sets of timestamped actions, namely user joining the group and photos uploaded in the goop pool. Therefore, similarly to the spatial clustering, we apply X-Means to this 8-dimensional feature space, obtaining three different clusters.

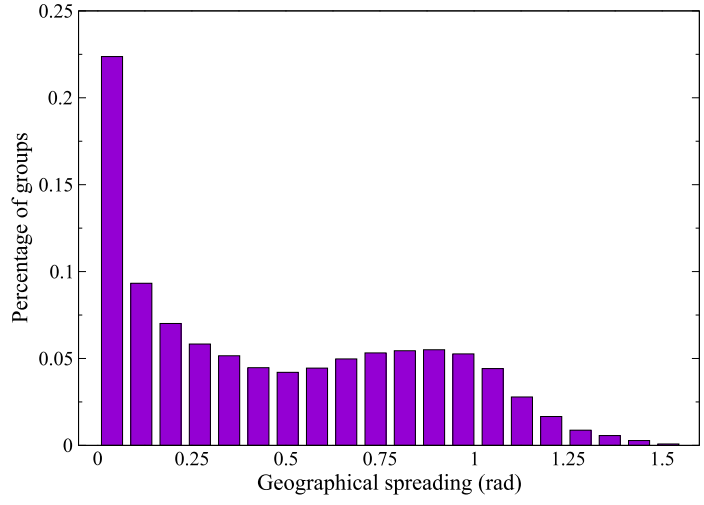


Figure 5 Histogram of the coe_d dispersion values of Flickr groups (values transformed in radians).

Table 2 Average and standard deviation of every feature in each of the clusters.

	Photos				Users			
	Cent.	Disp.	Skew.	Burst.	Cent.	Disp.	Skew.	Burst.
Evergreen	**0.48** ± **0.16**	0.56 ± 0.14	−0.01 ± 0.32	2.26 ± 1.85	**0.45** ± **0.14**	0.58 ± 0.13	−0.03 ± 0.29	0.72 ± 1.32
Short-lived	**0.03** ± **0.07**	**0.12** ± **0.17**	0.19 ± 0.47	1.82 ± 1.88	0.05 ± 0.09	**0.16** ± **0.18**	0.16 ± 0.54	0.62 ± 1.13
Bursty	**0.23** ± **0.15**	0.56 ± 0.19	0.43 ± 0.43	2.61 ± 1.92	0.15 ± 0.11	0.60 ± 0.19	**0.73** ± **0.32**	**2.30** ± **1.98**

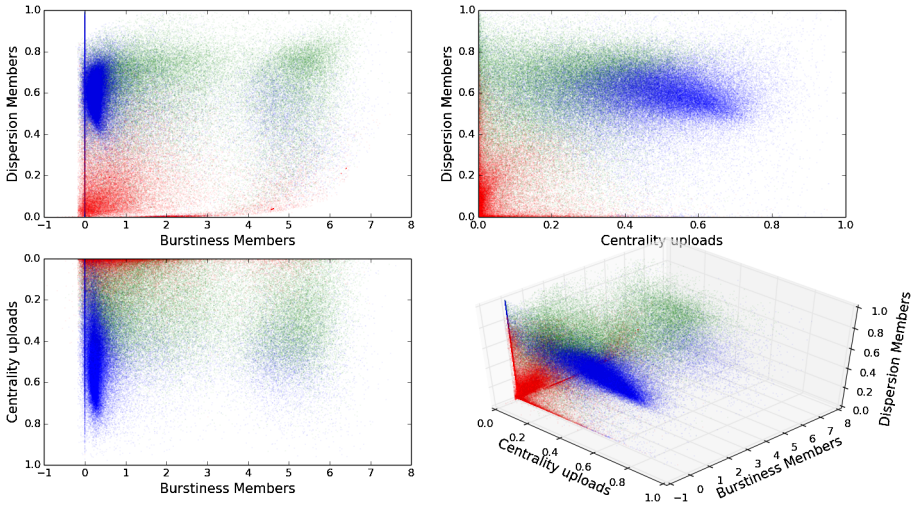


Figure 6 Scatter plot of the groups with respect to the three most discriminative features for the clustering on the temporal dimensions. Bursty groups are depicted in green, evergeen in blue, and short lived in red.

The average and standard deviation of every feature are displayed in Table 2. We find that the three most characteristic features are the dispersion and burstiness over users joining, and the centrality over photos uploaded. Figure 6 shows a scatter plot of these three features for each cluster. After an inspection of the clusters, we name them *evergreen*, *short-lived* and *bursty*. We report their peculiar features next.

Short-lived. The short-lived groups represent 13% of our sample and are characterized by low centrality and small dispersion. This category includes groups that had a little bit of activity after they were created and that became inactive shortly after. Examples include limited-scope photo sharing groups whose activity ceases shortly after the photos are uploaded and consumed by small social circles.

Evergreen. The evergreen cluster is the biggest one, containing 52% of the groups. Groups in this cluster are characterized by their high centrality and dispersion values around 0.5. were created at a certain point in the past and they have been growing in number of users and photos uniformly until the end of the time period we consider. Examples include groups dedicated to general topics, such as groups hosting artistic portraits from amateur and professional photographers.

Bursty. The remaining 34% of the groups are in the Bursty cluster, containing groups with lowest skewness and big burstiness, especially in the number of users joining. Those groups have usually the highest activity at the beginning of their life but then from time to time they experience photo uploads or user subscriptions in big batches. Some of these groups are related to recurring (e.g., yearly) events that attract attention of users regularly.

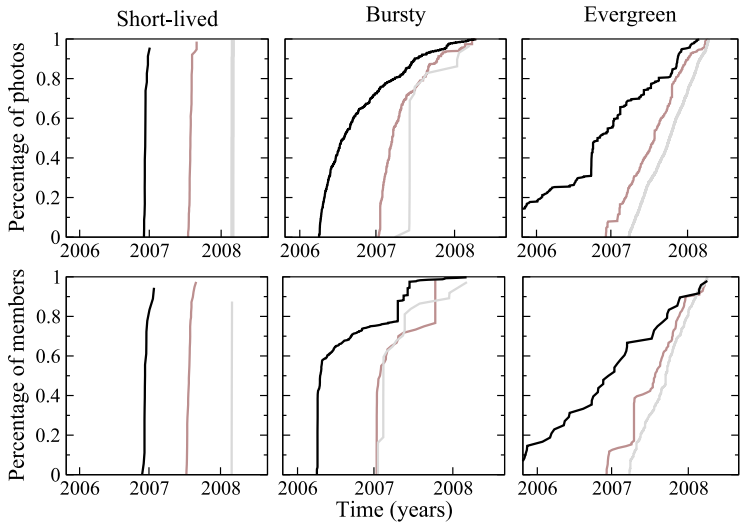


Figure 7 Evolution of each of the three most centered groups in each temporal cluster. Values on the *y*-axis are normalized by the maximum values reached at the end of the time window of our dataset.

The evolutions of number of users and photo uploads for the three most representative groups in each class are shown in Figure 7.

5.3 Socio-topical classes

To tackle the socio-topical dimension we first characterize the two sets of groups in terms of the metrics we introduced in Section 3.3.2. Then we study the relation between the labels of the declared groups annotated by the human coders and the values of the metrics. Additionally, we report ratios of groups labeled as social and topical among both declared, and detected groups.

5.3.1 Statistical properties of metrics

Besides directly comparing membership overlap, we study the variation of the metrics defined in Section 3.3.2 with the group size. Reciprocity and normalized entropy have a wide local maximum for groups of sizes between 50 and 100 members, both for declared and detected groups, as shown in Figures 8(a)-(d). This holds for all interactions and all sets of tags, with the exception of contacts, for which the curves are relatively flat. A similar result has been reported in a recent study [48] for pairwise interactions in Twitter by various community detection algorithms. We perform a random reshuffling of photos between groups, keeping the number of photos per group fixed. The normalized entropy calculated for the shuffled photos stays close to 1, as expected, and the maximum disappears. A possible interpretation of the existence of the maximum is that these sizes tend to correspond to social groups, while bigger groups are more frequently topical. Further findings to support this interpretation are presented in the next subsection.

Strong correspondence of the maxima for normalized entropy and reciprocity suggests that these properties are correlated, as shown in Figure 9. While it may be natural to explain the correlation between reciprocity of comments and normalized entropy based on commented photos, it is not clear why we also find a positive correlation with normalized entropy based on other sets of photos. A possible interpretation is that high intra-reciprocity leads to wider variety of topics covered inside of that group, and vice versa.

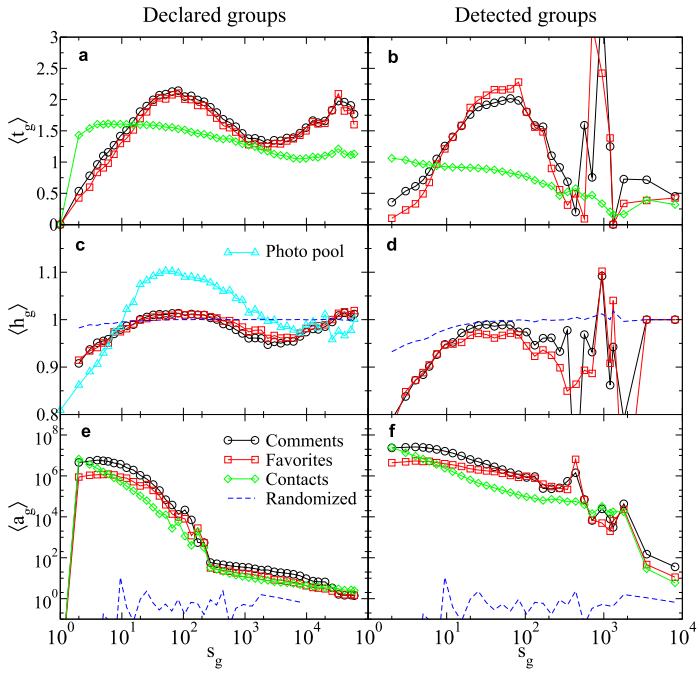


Figure 8 Fluctuation of metrics with group size. Dependency of normalized reciprocity **(a-b)**, normalized entropy **(c-d)** and relative activity **(e-f)** on size of groups for comments, favorites, contacts and photo pools, for declared and detected groups. Blue dashed line is for randomized photos **(c-d)** and groups **(e-f)**.

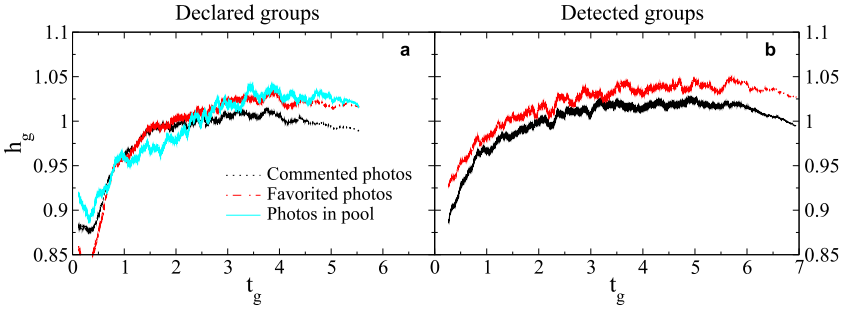


Figure 9 Correlation between reciprocity of comments inside a group and entropy of photos commented or favorited between its members, or belonging to the photo pool of this group, for declared and detected groups.

The values of relative activity both in declared and detected groups are very high, as presented in Figures 8(e), (f). As expected, activity of randomized groups exhibits values around 1 for all group sizes. For real groups instead, the value of relative activity decreases with the size of groups and gets close to 1 for very large ones. This may be caused by the fact that larger groups cannot be as integrated as smaller groups and the social commitment of their members towards other members of the group drops due to limited human capabilities. Additionally, we observe that the activity decay for declared groups occurs sharply between groups of size 100 and 200, in agreement with Dunbar's theory on the upper bound of the number of stable relationships manageable by a human. The activity drop for detected groups is continuous and much more moderate (Figure 8(f)), since com-

munity detection algorithms tend by design to output node clusters with high numbers of connections between them.

5.3.2 Relation between metrics and group label

Here we analyze properties and values of the metrics for groups labeled through the editorial process. First, the ratio of groups labeled as social differs between declared and detected groups. In declared groups we find around 48% social groups, whereas among detected groups almost 69% are labeled as social. Additionally, we picked 50 detected groups among the ones that are the most similar to declared groups. Specifically, we selected them randomly from the 99th percentile shown in Figure 4. These groups have significant overlap with declared groups and should share similar properties. Indeed, the ratio of groups labeled as social among them is closer to that of declared groups and equal to 53%. We conclude that detected groups are more likely to be social than declared ones. It is a somewhat expected result, since clustering algorithms detect dense parts of a network, and so they are inclined to detect areas with more reciprocal connections. Note that the theory envisions more reciprocal relations in social groups. Thus, community detection algorithms are more likely to find social groups, however, determining to what extent it happens is not trivial.

One of the expectations is that bond-based groups should not be very large, as the human capacity for stable relationships is limited. As pointed in Section 5.3.1, the Dunbar number can be considered as a possible cap for the size of such groups, while topical groups do no yield such a restriction. In line with this expectation, we find that declared groups labeled as social have on average 35 members, whereas groups labeled as topical have on average around 172 members.

We find insightful differences and similarities in various properties, which we explore in detail in Figure 10. We plot them as a function of the size of groups as they vary drastically with it, and one needs to compare groups of similar sizes in order to draw unbiased conclusions.

First, there are almost no differences in the number of photos (not shown), favorites, and contacts (as in Figures 10(b), (c)) inside social and topical groups. The number of comments is, however, around 2 times higher in social groups than in topical groups of similar size (Figure 10(a)). More differences can be found when looking at relative activity (Figures 10(d)-(i)), which compares the interaction internal to the group with the overall activity level of users belonging to groups. In all three types of interaction the relative activity metrics for social groups yield values from 2 to over 10 times higher than for topical groups. These metrics compare activity internal to the group with activity external to it. Therefore this result may reflect a stronger focus or even a possible isolation of members belonging to social groups from the rest of people they interact with.

More importantly, we observe large differences in values of reciprocity and relative reciprocity of comments and favorites. Social groups exhibit significantly higher reciprocity than topical groups (Figures 10(j)-(o)), in line with common identity and common bond theory. There is no difference in reciprocity of contacts, and a plausible interpretation is that contacts do not reflect personal relations between connected users. Possibly, since contacts do not need to be reciprocal, users often add people they do not know and do not interact with as contacts. Finally, we observe much higher values of entropy and normalized entropy in social groups than in topical ones (Figures 10(p), (q), (s), (t)). This holds

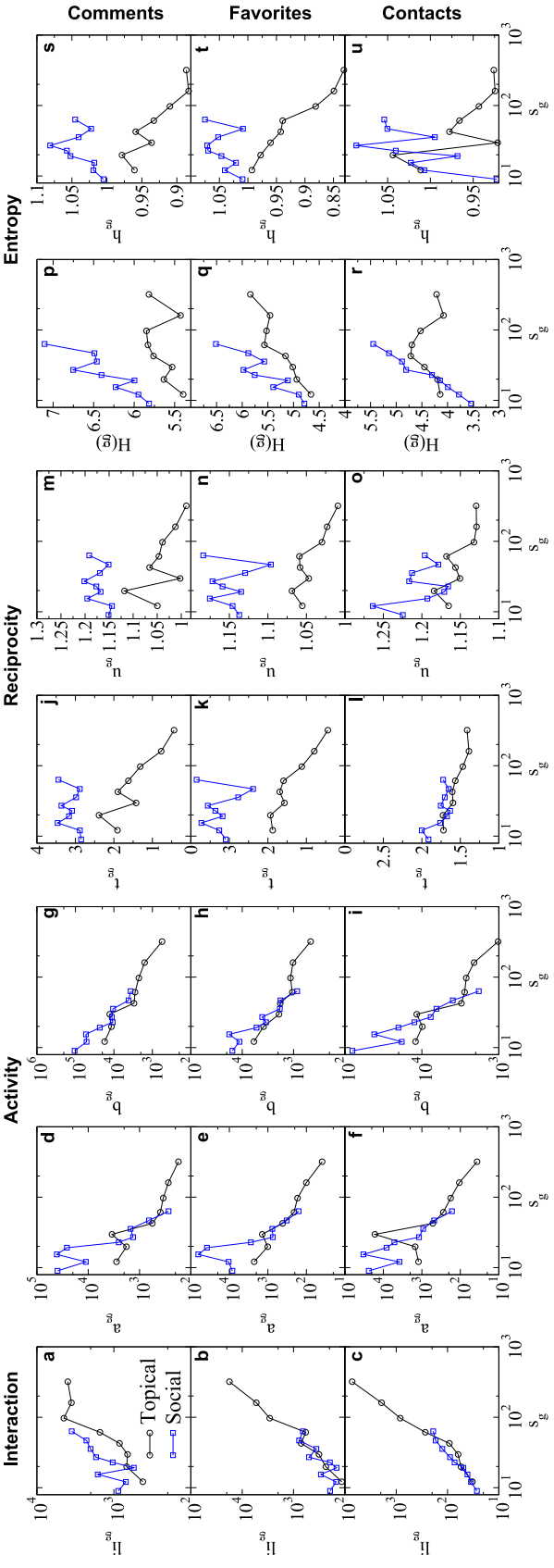


Figure 10 Averages of various topical (black circles) and social (blue squares) groups as a function of their size. Each point corresponds to 30 groups.

for the tags extracted from photos commented, and favorited between members. Assuming that tags of photos represent topics of interaction, the result is consistent with bond attachment. It is expected for members of bond-based groups to cover many different topics and areas in their interactions, whereas members of identity-based groups focus their interactions on specific topics. However, this does not hold for the tags extracted from photo pool of the group (Figures 10(r), (u)). Apparently, the content of the photo pool does not always reflect well the interactions and relations between members of the group.

Additionally, we plot the fraction of groups labeled as social with respect to group size, activity, reciprocity, and entropy (Figure 11). The size of the groups correlates negatively as expected (Figure 11(a)). The correlations with the number of interactions and relative activity a_g are rather weak (Figures 11(b), (c)), whereas surprisingly there is a strong dependency on relative activity b_g (Figure 11(d)). For the lowest values of b_g^{com}, 95% of the groups are topical, while for the highest, 80% of the groups are social. High values of b_g can mean stronger group-focus, or even an isolation of the group members from the rest of people they interact with. It may relate to the fact that it is hard to enter bond-based groups due to strong relations existing between their members and because high investment is required to create such relations with them [54]. Direct reciprocity of interactions, with the exception of contacts, correlates strongly with social groups (Figures 11(e), (f)). We strongly expected this result based on bond attachment. Furthermore, we found that the entropy of tags correlates with social groups, but entropy based on other sources does not (Figure 11(g)). However, we find that our normalized entropy performs much better than this, and a strong correlation is found both for tags extracted from comments and from favorites (Figure 11(h)). This shows that the normalized entropy of tags may be the most proper way of measuring topical diversity of communications of a set of people.

5.3.3 Group type detection

The properties of labeled social and topical groups tend to confirm the validity of the principles identified by the common identity and common bond theory. A stronger confirmation would directly come from the ability of the defined metrics to predict the tendency of a group towards sociality or topicality. To this end, we propose and compare two methods to predict the group type and we test their accuracy over the corpus of the labeled groups.

Prediction methodology The first approach we use is a linear combination of the metrics. To this end, we select the features that are the most related to the sociological theory and for which we built specific metrics, i.e., t_g, u_g and h_g. Each of them is applied to the 3 different interaction types and bags of tags, which produces a total of 9 values. We transform the values of the metrics into their t-statistics by subtracting the average value and dividing them by the standard deviation of the distribution. Then we weight the normalized scores evenly by dividing them by the total number of metrics considered and we finally sum them up to obtain a single *score* S_g. All of the components are supposed to score high for social groups. Therefore, the higher the value of the score, the higher the chance that the group is social rather than topical. To convert the score into a binary label, a fixed threshold above which groups are predicted to be social must be selected. Using this approach, we aim at testing if those metrics, based on the theory, can be successful in predicting the type of group (social or topical).

The second approach relies on machine-learning methods that use the metrics' values as features. Features are combined in a classifier that is first trained on a sample of labeled

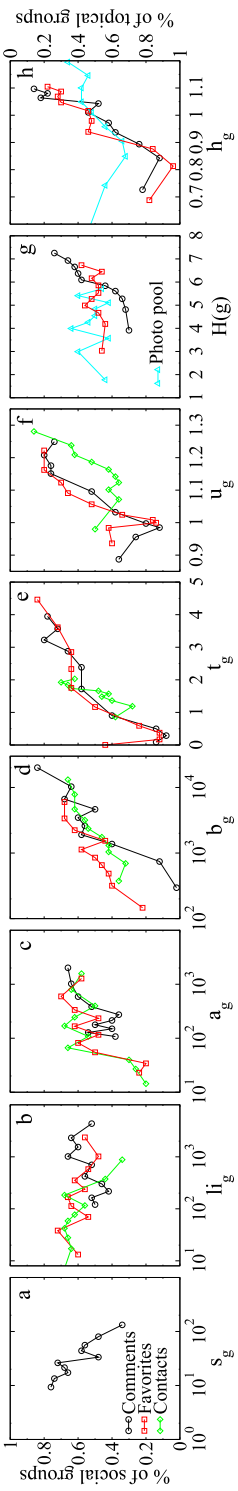


Figure 11 Dependence of fraction *f* of groups labeled as social on various metrics: based on comments, favorites, contacts and photo pools. The remaining (1 – *f*) groups are topical. Each point corresponds to 50 groups.

data to learn a prediction model. The trained classifier then outputs a binary prediction for any new group instance defined in the same feature space. Due to the limited size of our corpus of labeled groups, we estimate the classifier performance using 10-fold cross validation. We report results on a Rotation Forest classifier, which performed best in comparison to several algorithms implemented in WEKA. For the classifier we used a wider set of features than for the linear combination approach, namely: group size s_g and E_g^{int}, a_g, b_g, t_g, u_g, $H(g)$, h_g, each applied to the 3 different interaction types and bags of tags. This results in a total of 22 features. We selected such a wide set of features to test if indeed the metrics proposed to distinguish between the social and topical groups are the best ones for the task. The relative predictive power of the features is measured through a feature selection algorithm.

Prediction results The ratio of groups labeled as social increases quickly with the score S_g, as shown in Figure 12(a). This summarizes the findings of previous sections, suggesting that the features embedded in the score are able to capture well the nature of the groups. The higher the score, the higher the probability that the group is social; the lower, the more topical. When the score is around zero, groups can be either social or topical, or both, meaning that a decision on the nature of the group may be more difficult. If we fix the threshold for the S_g value in order to perform a binary group classification, it is clear that several misclassifications will occur, especially around the threshold value. An example for threshold at 0 is shown in Figure 12(a). Conversely, the classifier performs much better and achieves the ratio that adheres much more to the actual ratio of social and topical groups.

Both methods, however, fail more frequently for groups with mixed social and topical features. The prediction accuracies of the classifier and of the score-based predictions have an evident drop of performance around 0 (Figure 12(b)). The accuracy at the extreme values of the score is close to 0.95, while it falls below 0.6 for groups with a score close to 0. On the other hand, this drop appears also in the agreement between two of the human

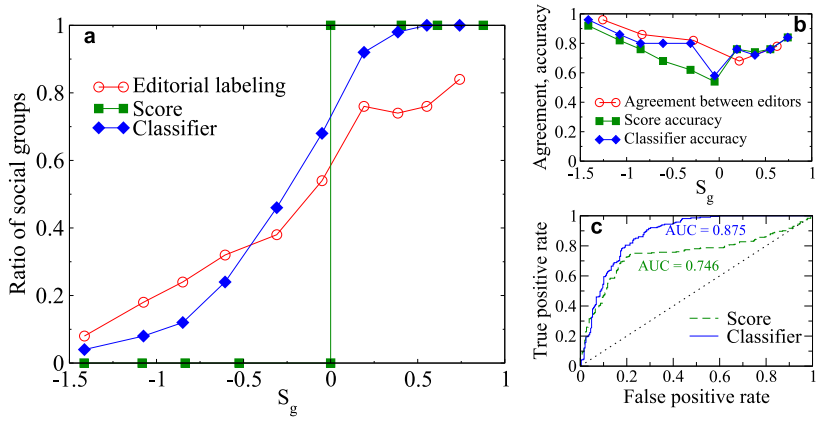


Figure 12 Comparison between prediction methods. (a) Ratio of groups labeled as social versus the score. The same ratio is plotted for the score-based (threshold at 0) and trained classifier predictions. **(b)** The accuracy of prediction of the two techniques and agreement between two of the labelers against the S_g values. **(c)** ROC curves for the prediction with the two different techniques.

Table 3 Group type prediction performance using (i) the score with threshold at 0, (ii) 10-fold cross validation on a Rotation Forest classifier trained on all the features, or (iii) the same classifier trained on the set of top-5 predictive features, according to the Chi Squared feature selection.

Method	Accuracy	AUC
Score	0.763	0.749
Classifier	0.801	0.879
Classifier χ^2_{top5}	0.803	0.872

labelers, measured as a ratio of groups that have been given the same label. Apparently, this is a shortcoming of the binary classification itself, as opposed to multi-label classification.

The overall performance of the two approaches can be compared fairly through ROC curves (Figure 12(c)), which astray from the selection of a fixed threshold. The curve for the classifier (computed for the 10-fold cross validation) always performs better, and this is reflected in the considerably higher AUC value and accuracy, as shown in Table 3.

In addition, to determine the most predictive features, we rank the features using Chi-square feature selection. The top 5 features are, in decreasing order of importance: h_g^{com}, t_g^{com}, u_g^{com}, h_g^{fav}, and b_g^{com}. The selected set is the optimal for the prediction performance: retraining the classifier on such restricted set of features results in stable performance, as shown in Table 3. The top 4 most predictive features correspond directly to the expectations of the theory and results of the analysis from Section 5; in other words, the normalized entropy of comments on the photo within the group and the reciprocity of comments exchanged between members are the best predictors of the socio-topical divide of groups. More surprisingly, as not explicitly mentioned in the original theory, also the amount of activity, namely the normalized activity in commenting in our case (b_g^{com}), is another good predictor. However, this is understandable, as we have already remarked on its importance and commented on its interpretation in Section 5.

5.4 Three-dimensional characterization

Once groups are characterized by each aspect separately, a natural question is whether there are some cross-dimensions relationships between group types, or in other words if different clusters of groups in one dimensions correspond predominantly to some other type of group in the other dimension. Blending all the metrics in a single model would be a way to answer the question. However, such unifying approach would be quite unpractical because of the different nature of the group characterization problem in different dimensions (clustering for geo-temporal, classification for socio-topical) and because of the difficult interpretation of a model that blends together such diverse types of measures.

For these reasons, we use a more modular and simple approach to analyze groups along the three dimensions together just by looking at the intersections between different classes. In this way we obtain an easier interpretation of results. Since there are two spatial classes (geo-wide and geo-narrow), three temporal classes (evergreen, short-lived and bursty), and two socio-topical classes (social and topical), there are 12 possible combinations of classes. The relative volume of the Flickr groups in our sample for each of them is reported in Table 4.

Some clear patterns emerge. First, social groups have a much higher ratio of bursty to evergreen groups than the topical ones. This is likely caused by the type of social behavior: a group of individuals who know each other would more likely join all the groups at its very

Table 4 Percentage of groups in each intersection between clusters. The sum of all the cells is 100%.

	Topical			Social		
	Short-lived	Evergreen	Bursty	Short-lived	Evergreen	Bursty
Geo-narrow	4.8%	15.8%	5.7%	5.3%	10.9%	12.7%
Geo-wide	1.4%	15.5%	4.2%	1.5%	9.7%	11.4%

beginning and probably would have a bursty activity caused by events of the social group. Symmetrically, topical groups tend to belong more to the "evergreen" category, as some topics are indeed not tied to the churn of social groups or to temporal trends. Furthermore, we can see a relation between short-lived and geo-narrow groups: groups that live for a short time have way less probability to spread on a big geographical scale, or in other words geo-width is a good indicator of an high chance of the group to survive longer.

6 Information diffusion in groups

Work in graph mining and social network analysis is too often conducted in several separate sub-branches focused on the solution of smaller tasks and with scarce contamination with other closely related pieces of research. One example is the relationship between communities and information diffusion. As cleverly noted just recently in a book by Easley and Kleinberg [71], the phenomenon of information diffusion, namely the flow of information along social links generating *information cascades* on a social network, is likely strongly coupled with the concept of community. In fact, the community boundaries should include people that are, to some extent, more similar to each other than to the rest of the network and the information likely would tend to spread inside that community and have lower penetration on the outside. In short: "cascades and clusters truly are natural opposites: clusters block the spread of cascades, and whenever a cascade comes to a stop, there's a cluster that can be used to explain why" [71].

Very recently, this idea has been leveraged by Barbieri *et al.* [7] who used data of information cascades to detect hidden communities. However, we argue that the process of spreading could be determined also by the type of communities involved in the process. Intuitively when a piece of information about a certain topic reaches a community that is interested in the same topic then the information will probably spread more easily, but what if a social (instead of topical) community is reached by the information cascade?

We contribute to shed light on this matter by running an experiment to check information cascades in relation to the types of groups we identify in this work. To do that, we rely on a well-established work by Cha *et al.* [61, 72] which modeled information propagation on Flickr. Here we replicate that model and study the resulting information cascade considering groups as additional component.

Cha *et al.*'s information diffusion model works using favorites. To consider a piece of information to propagate from user u_1 to user u_2 the following conditions must hold in a strict temporal order:

(1) u_2 starts following u_1;
(2) u_1 favorites a photo p;
(3) u_2 favorites the same photo p.

This experimental framework is motivated by the fact that, in Flickr, users are notified about the photos that their followees favorite. The information diffusion links can be used

to reconstruct potentially several information diffusion cascades (also called "diffusion trees"), where the *root* is a user who favorited a photo without having any followees who favorited it before.

To check the relation between cascades and group types, we have to extend the aforementioned framework by embedding the notion of group. Specifically, we want to check whether a photo that is uploaded in a group pool has a diffusion that is predominantly restricted to that group or spreads beyond the group boundaries. Therefore, we consider roots of our diffusion trees all the users that comply with the following strict temporal sequence:

(1) user u joins group g;
(2) photo p is uploaded to g;
(3) u favorites p.

Of course, for each (g,p), $g \in G \wedge p \in P$ pair there could be multiple root users, namely multiple members of the group who are not following each other and who all favorite the same photo according to the temporal sequence specified above. We connect all this root users to a common super-root identified by the (g,p) pair. Once the root nodes are identified, we apply the framework by Cha *et al.*, thus obtaining information cascades each labeled by a unique (g,p) pair. Note that a photo could be uploaded in multiple group pools, thus originating more than one cascade. We consider each of these possible cascades separately.

The method we propose is limited by the fact that the root user might favorite a photo not because it has been published in a group but for any other reason (e.g., it was discovered by random browsing). However, we argue that if the photo has been uploaded to the pool we can assume it to be relevant to the group and the nature of the actual action that triggered the first favorite is not crucial to the study.

Given this setting, for each cascade we compute a pair of values. Considering $A_{g,p}$ to be the set of adopters, namely the users who take part in the diffusion tree for the (g,p) pair, and M_g the members of group g, we define:

$$c_{g,p} = \frac{|A_{g,p} \cap M_g|}{|M_g|}, \tag{12}$$

$$s_{g,p} = 1 - \frac{|A_{g,p} \cap M_g|}{|A_{g,p}|}. \tag{13}$$

The $c_{g,p}$ measure the coverage of the cascade, namely the reaction to a photo by the community or, in other words, how much the photo spread inside the group. The $s_{g,p}$ measures instead the *external spreading* and captures how much the information spreads outside group. An example of a cascade is given in Figure 13.

To characterize each group, all the values $c_{g,p}$ and $s_{g,p}$ are averaged for all their photos, leading to the aggregate values c_g and s_g. To study how the information spreads in groups of different types, we consider the values for each of the group types separately and we compute the average values at a fixed group size, to account for any dimensionality effects. The results are shown in Figures 14, 15, and 16.

On the social-topical axis, the difference between different type of groups is slight but noticeable, with the topical groups having slightly more coverage and the social groups more external spreading (except for a small range of group sizes). This supports the intuitions of previous work that identifies the boundaries of topical groups as harder to be

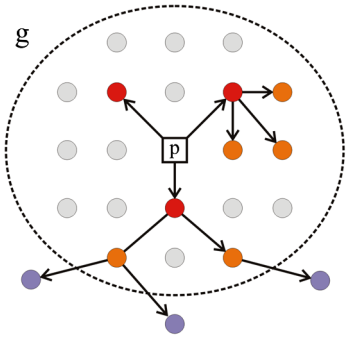


Figure 13 Example of diffusion tree for a photo *p* uploaded into the photo pool of group *g*. Circles represent users and the dashed line marks the boundaries of the group. Red circles are the root users. In this example 8 users out of 20 members are nodes of the diffusion tree, leading to $\text{coverage}_{g,p} = 0.4$. Also, 3 users outside the group are nodes of the tree, for a total tree size of 11 nodes (except the meta-root), thus leading to $\text{externalCoverage}_{g,p} = 0.27$.

crossed by information cascades. This is somehow expected in the case in which members of topical groups share interests which are narrow enough to be limited predominantly to the groups members, while members of social groups do not necessarily share a specific common interest, therefore their favoriting behaviour is more varied and with higher chance to have an echo also outside the group. On the geographical dimension instead the difference is almost negligible, with slightly higher values for geo-wide groups for both metrics. This might be related to a better capacity of geo-wide groups to spread information in general.

More evident trends are obtained on the time dimension. On average, the evergreen groups have more coverage than the short-lived or the bursty ones, whereas the bursty groups are the ones with most external spreading. Evergreen groups are always active, so they get a lot of attention from their members, partially explaining why photos published in them get more coverage. On the other hand, bursty groups are often related to major events with broad scope whose photos can be of interest to a large audience in the Flickr community not restricted to the members only.

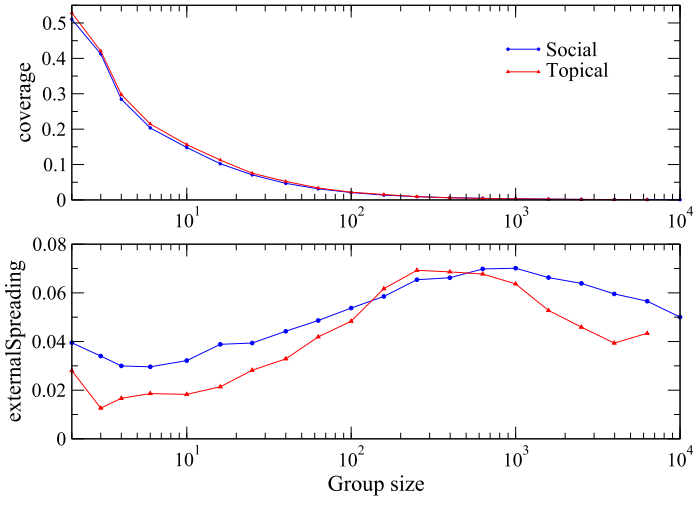


Figure 14 Average values of coverage and external spreading for all groups with same number of members. Social and topical group types are considered.

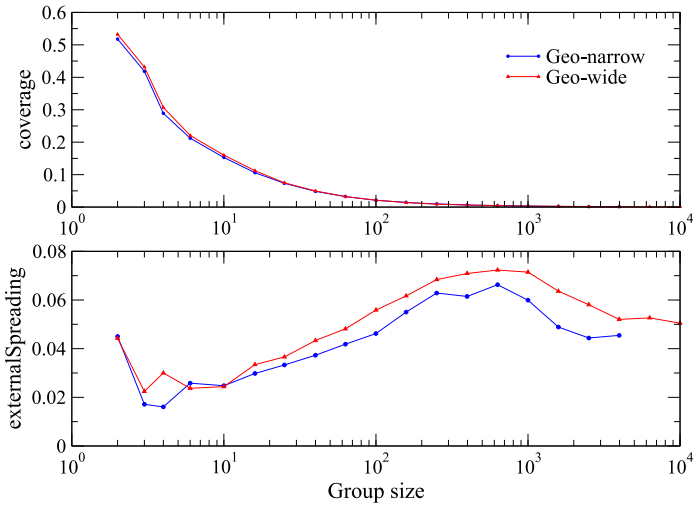


Figure 15 Average values of coverage and external spreading for all groups with same number of members. Geo-narrow and geo-wide group types are considered.

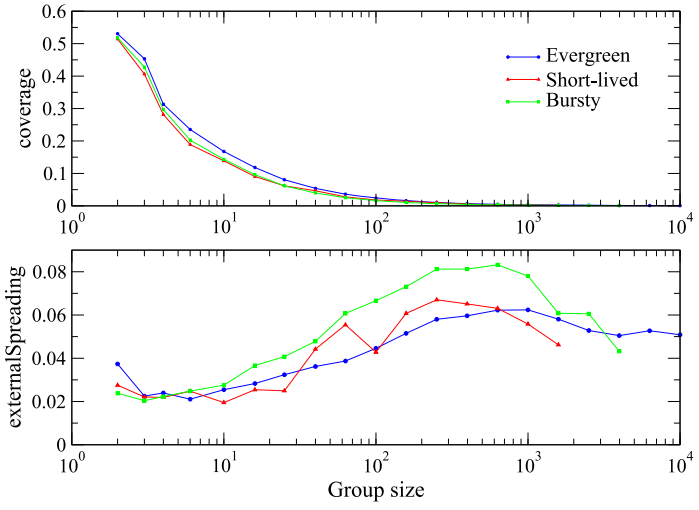


Figure 16 Average values of coverage and external spreading for all groups with same number of members. Evergreen, short-lived, and bursty group types are considered.

7 Conclusions

Providing nuanced descriptions of interaction atoms in networked social systems is crucial to get an accurate understanding of the online collective human behaviour. After social links, groups are the most important social structures around which the activity of social media revolves.

We contribute to explore this area by proposing a set of general metrics to capture the spatial, temporal, and socio-topical dimensions of groups, which are the three aspects about groups that have been informally identified in the previous literature but never formalized and studied in conjunction. Using a large Flickr group dataset for our experiments, we propose a new metric to account for geographical sparsity that identifies two main classes of spatially-characterized groups (geo-narrow and geo-wide); we cluster groups

according to their temporal activity, being able to discover three major temporal patterns (evergreen, bursty, and short-lived groups); last we translate the "common identity and common bond theory" into metrics of reciprocity, activity, and topical diversity to distinguish social and topical groups. In particular, we annotate a number of Flickr groups as either topical or social and we match this ground truth with the machine-generated labels, showing that the socio-topical metrics, combined with a machine-learning approach, predict the group type with high accuracy. The analysis of the three dimensions in combination allows us to show interesting correlations between different classes. In particular, we find that groups that manage to spread on geographically-large scale are usually more long-lived than "local" groups, that topical groups tend to have a constant activity behaviour, being tolerant to the churn of their users, and that social groups have a bursty activity traces, with all the members joining at first and then interacting with each other from time to time, after relatively long periods of inactivity.

Besides these main results, our study is enriched by several pieces of complementary analysis. First, we find that the dependency of the socio-topical metrics on the group size confirms previous observations about the effective size of social communities (also known as Dunbar's Number), peaking around rather small sizes and being limited by a cap of 100-200 members. Also, the comparison of the structure and sociality and topicality traits between declared groups and groups from community detection algorithms reveals that detected groups do not overlap much with declared groups on average, but they match sensibly more than the random case for groups of comparable sizes. Furthermore, detected groups are more often social than the declared ones. Last, inspired by previous work that puts in relation communities and information cascades and relying on a well-established model of information diffusion on Flickr, we study the dependency between group type and volume of information spreading inside or outside a group. We find that social and bursty groups allow the information to spread across the boundaries of groups more than topical and evergreen groups, that instead tend to retain more the information within them.

We hope that our study brings a constructive message in terms of (i) the need of more nuanced description of the structures in social networks and (ii) the benefits of putting in relation different collective phenomena that are rarely put in relation one with another.

Appendix

A.1 Correction parameter for standard deviation

Standard formulation of standard deviation is:

$$\sigma^2 = \sqrt{\frac{1}{N-1}\sum(t-\mu)^2}. \tag{14}$$

Given a list N values t that can be assumed in $[0,1]$, with a given mean μ the greater possible standard deviation would be achieved under a Bernoulli distribution with $t = 1$ with probability p and $t = 0$ with probability q. Under these circumstances we can write:

$$\sum(t-\mu)^2 = N \cdot p \cdot (0-\mu)^2 + N \cdot q \cdot (1-\mu)^2, \tag{15}$$

which, under a Bernoulli distribution, can be rewritten as:

$$\sum (t-\mu)^2 = N \cdot (1-\mu) \cdot (0-\mu)^2 + N \cdot \mu \cdot (1-\mu)^2 \tag{16}$$

$$= N \cdot (1-\mu)\mu^2 + N \cdot \mu \left(1 + \mu^2 - 2\mu\right) \tag{17}$$

$$= N\mu(1-\mu). \tag{18}$$

Therefore, being $N\mu(1-\mu)$ the maximum value for $\sum(t-\mu)^2$, we use it as normalization factor in (2).

A.2 Correction parameter for skewness

Under a Bernoulli distribution that assumes value 0 with probability $p(0)$ and 1 with $p(1)$, the mean μ is equal to $p(1)$, while the median is given by:

$$\text{median} = \begin{cases} 0 & \text{if } p(0) > p(1), \\ 0.5 & \text{if } p(0) = p(1), \\ 1 & \text{if } p(0) < p(1). \end{cases} \tag{19}$$

In case $p(0) = p(1) = 0.5$ the normalization factor is not relevant so mean and median are equal and the difference would remain the same. In other cases, one can define the maximum difference ($\max_{\text{diff}}$) given the mean μ as follows:

$$\max_{\text{diff}} = \begin{cases} 1-\mu & \text{if } p(0) < p(1), \\ \mu & \text{if } p(0) > p(1). \end{cases} \tag{20}$$

Since the mean μ is equal to $p(1)$ and $p(0)$ is equal to $1-\mu$, we can rewrite the equation as:

$$\max_{\text{diff}} = \begin{cases} 1-\mu & \text{if } 1-\mu < \mu, \\ \mu & \text{if } 1-\mu > \mu, \end{cases} \tag{21}$$

that can be finally rewritten as:

$$\max_{\text{diff}} = \min(1-\mu, \mu), \tag{22}$$

which we use it as normalization factor in (3).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
DMB, LMA and PG designed the methodology, conceived and run the experiments. All authors wrote and revised the manuscript. This work was carried out while DMB and PG were interns at Yahoo Labs, Barcelona.

Author details
[1]Universitat Pompeu Fabra, Barcelona, Spain. [2]Yahoo Labs, Barcelona, Spain. [3]Max Planck Institute for Software Systems, Saarbruecken, Germany.

Acknowledgements
This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975.

Received: 5 March 2014 Accepted: 18 July 2014 Published online: 24 September 2014

References

1. Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. IMC'07. ACM, San Diego, pp 29-42
2. Negoescu RA, Gatica-Perez D (2008) Analyzing Flickr groups. In: Proceedings of the 2008 international conference on content-based image and video retrieval. CIVR'08. ACM, New York, pp 417-426
3. Kairam SR, Wang DJ, Leskovec J (2012) The life and death of online groups: predicting group growth and longevity. In: Proceedings of the fifth ACM international conference on Web search and data mining. WSDM'12. ACM, New York, pp 673-682
4. Aiello LM, Barrat A, Schifanella R, Cattuto C, Markines B, Menczer F (2012) Friendship prediction and homophily in social media. ACM Trans Web 6(2):9
5. Monge P, Contractor NS (2003) Theories of communication networks. Oxford University Press, London
6. Aiello LM, Schifanella R, State B (2014) Reading the source code of social ties. In: Conference on web science (WebSci'14). ACM, New York, pp 139-148. doi:10.1145/2615569.2615672
7. Barbieri N, Bonchi F, Manco G (2013) Cascade-based community detection. In: Proceedings of the sixth ACM international conference on Web search and data mining. WSDM'13. ACM, New York, pp 33-42
8. Grabowicz PA, Aiello LM, Eguiluz VM, Jaimes A (2013) Distinguishing topical and social groups based on common identity and bond theory. In: Proceedings of the sixth ACM international conference on Web search and data mining. WSDM'13. ACM, New York, pp 627-636
9. Dunbar RIM (1998) The social brain hypothesis. Evol Anthropol 6:178-190
10. Porter CE (2004) A typology of virtual communities: a multi-disciplinary foundation for future research. J Comput-Mediat Commun. doi:10.1111/j.1083-6101.2004.tb00228.x
11. De Choudhury M (2009) Modeling and predicting group activity over time in online social media. In: Proceedings of the 20th ACM conference on hypertext and hypermedia. HT'09. ACM, New York, pp 349-350
12. Wang J, Zhao Z, Zhou J, Wang H, Cui B, Qi G (2012) Recommending Flickr groups with social topic model. Inf Retr 15(3-4):278-295
13. Cox A, Clough P, Siersdorfer S (2011) Developing metrics to characterize Flickr groups. J Am Soc Inf Sci Technol 62:493-506
14. Grabowicz PA, Eguíluz VM (2012) Heterogeneity shapes groups growth in social online communities. Europhys Lett 97(2):28002
15. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. KDD'06. ACM, New York, p 44
16. Baldassarri A, Barrat A, Capocci A, Halpin H, Lehner U, Ramasco J, Robu V, Taraborelli D (2008) The Berners-Lee hypothesis: power laws and group structure in Flickr. In: Alani H, Staab S, Stumme G (eds) Social Web communities. Dagstuhl seminar proceedings. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Germany, Dagstuhl
17. Van House NA (2007) Flickr and public image-sharing: distant closeness and photo exhibition. In: Extended abstracts on human factors in computing systems. CHI'07. ACM, New York, pp 2717-2722
18. Miller AD, Edwards WK (2007) Give and take: a study of consumer photo-sharing culture and practice. In: Proceedings of the SIGCHI conference on human factors in computing systems. CHI'07. ACM, New York, pp 347-356
19. Nov O, Naaman M, Ye C (2010) Analysis of participation in an online photo-sharing community: a multidimensional perspective. J Am Soc Inf Sci Technol 61(3):555-566
20. Negoescu R-A, Adams B, Phung D, Venkatesh S, Gatica-Perez D (2009) Flickr hypergroups. In: Proceedings of the 17th ACM international conference on multimedia. MM'09. ACM, New York, pp 813-816
21. Negi S, Chaudhury S (2012) Finding subgroups in a Flickr group. In: Proceedings of the 2012 IEEE international conference on multimedia and expo. ICME'12. IEEE Computer Society, Washington, pp 675-680
22. Negoescu RA, Gatica-Perez D (2008) Topickr: Flickr groups and users reloaded. In: Proceedings of the 16th ACM international conference on multimedia. MM'08. ACM, New York, pp 857-860
23. Negoescu R-A, Gatica-Perez D (2010) Modeling Flickr communities through probabilistic topic-based analysis. IEEE Trans Multimed 12(5):399-416
24. Butler B (1999) When a group is not a group: an empirical examination of metaphors for online social structure. PhD thesis, Carnegie Mellon University
25. Holmes P, Cox AM (2011) Every group carries the flavour of the admins. Leadership on Flickr. Int J Web Based Communities 7(3):376-391
26. Prieur C, Pissard N, Beuscart J, Cardon D (2008) Thematic and social indicators for Flickr groups. In: Proceedings of ICWSM
27. Prieur C, Cardon D, Beuscart J-S, Pissard N, Pons P (2008) The strength of weak cooperation: a case study on Flickr. arXiv:0802.2317
28. Pissard N, Prieur C (2007) Thematic vs. social networks in Web 2.0 communities: a case study on Flickr groups. In: Algotel conference
29. Backstrom L, Kumar R, Marlow C, Novak J, Tomkins A (2008) Preferential behavior in online groups. In: Proceedings of the 2008 international conference on Web search and data mining. WSDM'08. ACM, Palo Alto, pp 117-128
30. Welser HT, Gleave E, Fisher D, Smith M (2007) Visualizing the signatures of social roles in online discussion groups. J Soc Struct 8:2
31. Gloor PA, Zhao Y (2006) Analyzing actors and their discussion topics by semantic social network analysis. In: Proceedings of the conference on information visualization. IV'06. IEEE Computer Society, Washington, pp 130-135

32. Spertus E, Sahami M, Buyukkokten O (2005) Evaluating similarity measures: a large-scale study in the Orkut social network. In: Proceedings of the 11th ACM SIGKDD international conference on knowledge discovery in data mining. KDD'05. ACM, New York, pp 678-684
33. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. KDD'06. ACM, New York, pp 44-54
34. Laine MSS, Ercal G, Luo B (2011) User groups in social networks: an experimental study on Youtube. In: 2011 44th Hawaii international conference on system sciences (HICSS), pp 1-10
35. Palla G, Barabási A-L, Vicsek T (2007) Quantifying social group evolution. Nature 446:664-667
36. Onnela J-P, Arbesman S, González MC, Barabási A-L, Christakis NA (2011) Geographic constraints on social network groups. PLoS ONE 6(4):e16939
37. Barabási A-L (2005) The origin of bursts and heavy tails in human dynamics. Nature 435:207-211
38. Riger S, Lavrakas PJ (1981) Community ties: patterns of attachment and social interaction in urban neighborhoods. Am J Community Psychol 9:55-66. doi:10.1007/BF00896360
39. Tajfel H (1982) Social identity and intergroup relations. Cambridge University Press, Cambridge
40. McGrath JE, Arrow H, Berdahl JL (2000) The study of groups: past, present, and future. Personal Soc Psychol Rev 4(1):95-105
41. McMillan DW, Chavis DM (1986) Sense of community: a definition and theory. J Community Psychol 14(1):6-23
42. Siegel J, Dubrovsky V, Kiesler S, McGuire TW (1986) Group processes in computer-mediated communication. Organ Behav Hum Decis Process 37(2):157-187
43. Walther JB (1997) Group and interpersonal effects in international computer-mediated collaboration. Hum Commun Res 23(3):342-369
44. Saunders CS, Ahuja MK (2006) Are all distributed teams the same? Differentiating between temporary and ongoing distributed teams. Small Group Res 37(6):662-700
45. Wang X, Butler BS, Ren Y (2013) The impact of membership overlap on growth: an ecological competition view of online groups. Organ Sci 24(2):414-431
46. Butler B, Sproull L, Kiesler S, Kraut R (2008) Community effort in online groups: who does the work and why? In: Leadership at a distance
47. Postmes T, Spears R, Lea M (2000) The formation of group norms in computer-mediated communication. Hum Commun Res 26(3):341-371
48. Grabowicz PA, Ramasco JJ, Moro E, Pujol JM, Eguiluz VM (2012) Social features of online networks: the strength of intermediary ties in online social media. PLoS ONE 7(1):e29358
49. Goncalves B, Perra N, Vespignani A (2011) Modeling users' activity on Twitter networks: validation of Dunbar's number. PLoS ONE 6(8):e22656
50. Tang L, Wang X, Liu H (2011) Group profiling for understanding social structures. ACM Trans Intell Syst Technol 3(1):15
51. Ludford PJ, Cosley D, Frankowski D, Terveen L (2004) Think different: increasing online community participation using uniqueness and group dissimilarity. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 631-638
52. Prentice DA, Miller DT, Lightdale JR (1994) Asymmetries in attachments to groups and to their members: distinguishing between common-identity and common-bond groups. Pers Soc Psychol Bull 20(5):484-493
53. Sassenberg K (2002) Common bond and common identity groups on the Internet: attachment and normative behavior in on-topic and off-topic chats. Group Dyn 6(1):27-37
54. Ren Y, Kraut R, Kiesler S (2007) Applying common identity and bond theory to design of online communities. Organ Stud 28(3):377-408
55. Fortunato S (2010) Community detection in graphs. Phys Rep 486(3-5):75-174
56. Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78:046110
57. Yang J, Leskovec J (2012) Defining and evaluating network communities based on ground-truth. arXiv:1205.6233
58. Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining. KDD'03. ACM, New York, pp 137-146
59. Ye S, Wu SF (2010) Measuring message propagation and social influence on twitter.com. In: Proceedings of the second international conference on social informatics. SocInfo'10. Springer, Berlin, pp 216-231
60. Cha M, Haddadi H, Benevenuto F, Gummadi KP (2010) Measuring user influence in Twitter: the million follower fallacy. In: 4th international AAAI conference on Weblogs and social media (ICWSM)
61. Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the Flickr social network. In: Proceedings of the 18th international conference on World Wide Web. WWW'09. ACM, Madrid, pp 721-730
62. Yang J, Leskovec J (2010) Modeling information diffusion in implicit networks. In: Proceedings of the 2010 IEEE international conference on data mining. ICDM'10. IEEE Computer Society, Washington, pp 599-608
63. Au Yeung C-m, Iwata T (2010) Capturing implicit user influence in online social sharing. In: Proceedings of the 21st ACM conference on hypertext and hypermedia. HT'10. ACM, New York, pp 245-254
64. Gomez Rodriguez M, Leskovec J, Krause A (2010) Inferring networks of diffusion and influence. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. KDD'10. ACM, New York, pp 1019-1028
65. Barbieri N, Bonchi F, Manco G (2013) Influence-based network-oblivious community detection. In: 2013 IEEE 13th international conference on data mining (ICDM), pp 955-960
66. Zwol RV (2007) Flickr: who is looking? In: IEEE/WIC/ACM international conference on Web intelligence. WI'07. IEEE Computer Society, Washington, pp 184-190
67. Utz S, Sassenberg K (2002) Distributive justice in common-bond and common-identity groups. Group Process Intergroup Relat 5(2):151-162
68. Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PLoS ONE 6(4):e18961

69. Collins NL, Miller LC (1994) Self-disclosure and liking: a meta-analytic review. Psychol Bull 166(3):457-475
70. Pelleg D, Moore AW (2000) *X*-means: extending *K*-means with efficient estimation of the number of clusters. In: Proceedings of the seventeenth international conference on machine learning. ICML'00. Morgan Kaufmann, San Francisco, pp 727-734
71. Easley D, Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, New York
72. Cha M, Mislove A, Adams B, Gummadi KP (2008) Characterizing social cascades in Flickr. In: Proceedings of the first workshop on online social networks. WOSP'08. ACM, Seattle, pp 13-18

doi:10.1140/epjds/s13688-014-0008-y
Cite this article as: Martin-Borregon et al.: **Characterization of online groups along space, time, and social dimensions.** *EPJ Data Science* 2014 **2014**:8.