Chapter 5
Group Types in Social Media

Luca Maria Aiello

Abstract Dynamics of social systems are the result of the complex superposition of interactions taking place at different scales, ranging from the pairwise communications between individuals to the macroscopic evolutionary patterns of the full interaction graph. Social communities, namely groups of people originated by any spontaneous aggregation process, constitute the mid-ground between such two extremes. Groups are important constituents of social environments as they form the basis for people's participation and engagement beyond their minute dyadic interactions. Communities in online social media have been studied widely in their static and evolutionary aspects, but only recently some attention has been devoted to the exploration of their *nature*. Besides the characterization of online communities along their spatio-temporal and activity features, the recent advancements in the emerging field of computational sociology have provided a new lens to study social aggregations along their social and topical dimensions. Using the online photo sharing community Flickr as a main running example, we survey some techniques that have been used to get a multi-faceted description of group types and we show that different types of groups impact on orthogonal interaction processes on the social graph, such as the diffusion of information along social ties. Our overview supports the intuition that a more nuanced description of groups could not only improve the understanding of the activity of the user base but can also foster a better interpretation of other phenomena occurring on social graphs.

5.1 Bridging Gaps in the Study of Communities

> Most human pleasures have their roots in social life. [...] Much of human suffering as well as much of human happiness has its source in the actions of other human beings. One follows from the other, given the facts of **group** life, where pairs do not exist in complete isolation from other social relations.

L.M. Aiello (✉)
Yahoo Labs, 125 Shaftesbury Avenue, London, UK
e-mail: alucca@yahoo-inc.com

G. Paliouras et al. (eds.), *User Community Discovery*, Human–Computer Interaction Series, DOI 10.1007/978-3-319-23835-7_5

This is how sociologist Peter Blau introduces the discussion about the structure of social associations in his famous book "Exchange and Power in Social Life" [11], acknowledging the pivotal role of groups in providing motivations and rewards for people in a social ecosystem. Together with dyadic social interactions, groups form the basis for the social instantiation of any individual.

Given the centrality and pervasiveness of such social structures in our everyday life, it is no surprise that their transposition in online social media has gained an explosive and apparently ever-growing success. Besides providing the possibility to establish pairwise social connections online, social media allow for the creation of groups (or communities[1]) that are characterized, depending on the online system considered, by different properties [30, 40, 44]. As a result, groups in social media have flourished and they nowadays form a strong basis for user participation and engagement in online services.

Groups, either online or offline, have been the object of studies in social sciences for decades and yet, because the notion of group itself hides an enormous variety of concepts representing as many group types in real life, it is very difficult to provide a general definition of what a group really is. In fact, a group can be characterized simply by a social aggregation involving more than a certain number of actors, as well as by more abstract concepts such as similarity or interdependence of the members. One of the most well-established interpretation of the meaning of groups is based on the notion of *social identity*, an elusive idea that is hard to frame and has been object of debate and investigation. Social identity is understood by the social psychologist Henri Tajfel, one of the pioneers of the social identity theory, as the part of an individual's self-concept deriving from the membership of a social group, together with the emotional valuation that the membership may imply [56]. Tajfel has himself acknowledged that the discussion on what identity is can be often "endless and sterile" [57] because of the complexity of social interactions that surround an individual.

On a parallel track, computer science research has partly confirmed some of the key notions illustrated above through data-driven studies. By intensively investigating online groups, evidence has been found about the tendency of actors to flock in communities pushed by a number of reasons including affiliation by similarity, common interest, conflict with other groups, local proximity, or even just by the need of defining a distinctive identity with respect to the rest of the population [1–3, 31, 38, 40].

Despite all the efforts spent in the study of online groups, there are still some major gaps that just recently have begun to be filled to reach a more coherent, complete and nuanced description of the nature of groups. First, the research community has mainly considered groups as homogeneous entities, overlooking the fact that groups are not all created equal, as they emerge by different collective processes and by the

[1] The distinction between "group" and "community" is very subtle and varies in different research fields. If not specified differently, we will use the two terms interchangeably in this chapter.

different motivations of their founders or members. Such lapse has been exasperated by the tendency of studying different characterizing dimensions of online groups—temporal, structural, spatial, etc.—in separation. Last, although computer science research on online groups has corroborated the theories developed in social sciences, more systematic approaches to the verification of sociological findings with computational methods have been emerging only very recently [18]. As a consequence, a thorough exploration of the *nature* of online groups is still a work in progress, with still very few systematic approaches to characterize groups along multiple quantitative dimensions.

This chapter aims to present recent work that has been directed to address the limitations mentioned above. We will describe work that has contributed to compose the fractiousness within the computer science literature by attempting multifaceted characterizations of groups. The work we describe also attempts to bridge the gap with social science studies by operationalizing theories about communities that have been previously developed in sociology.

Specifically, we will describe a categorization of online groups that considers *spatial*, *temporal* (Sect. 5.4) and *socio-topical* (Sect. 5.5) dimensions for the first time in combination (Sect. 5.6) and that captures in a computational framework an instantiation of the notion of common identity (as opposed to common bond) that has been for long time discussed in social sciences. We explore also the implications of the group size on its activity—in relation with the so-called Dunbar number theory (Sect. 5.8)—and discuss the relationship between communities that are spontaneously created by the user base and those that are algorithmically found by community detection algorithms, based on the density of interactions between actors (Sect. 5.7). Also, to further support the belief that a nuanced characterization of groups matters, especially if informed by notions coming from the social sciences, we speculate about the impact that different group types may have in another important social process occurring in social networks: information diffusion. We follow the intuition that the shape information cascades is partly determined by the type of community in which the piece of information is propagating (Sect. 5.9). Finally, we conclude by briefly discussing the role of social groups in addressing the micro-macro problem in sociology (Sect. 5.10).

Along the remainder of the chapter, our main case-study will be Flickr, the world-famous photo-sharing social platform. The experiments we report have ben run on a large scale Flickr dataset described in Sect. 5.3. Flickr has a rich set of features accessible via public API[2] including a direct social network, explicit declaration of groups, annotated content, dyadic conversations, etc. thus being an ideal dataset to explore different facets of social aggregations.

[2] https://www.flickr.com/services/api/.

5.2 Group Characterization in the Literature

Next, to provide a general context of the state of the art on the study of online and offline groups, we review some of the most notable work in the fields of computer science and social sciences that have been published around the topic. That will set the background upon which the following discussion will build on.

5.2.1 Groups in Computer Science

Social online communities have been investigated since the beginning of the social web. Besides spending efforts in finding empirical evidences to support different notions of communities [48], researchers have explored groups in relation to several applications including recommendation and profiling [19, 65]. The static structure, as well as the evolutionary dynamics of communities have been investigated extensively over a variety of large-scale and heterogeneous datasets. Extensive evidence has been produced about the broad distribution of the structural and temporal features of online social groups [16, 38]. Those characteristics are largely determined by the intrinsic group fitness [25] and by the density of social ties connecting their members [6].

Groups are extremely varied in terms of their emerging features, from their size [9] to their purpose [27]. Such variety has triggered a line of research work that attempted to capture the nature of social groups along several axes, but most often with a lack of any quantitative framework for their classification. Consequently, the results achieved in this area are mostly scattered and lack of consistence.

Due to its open nature, Flickr has been the most studied platform in this respect. Early work identified differences in the usage of Flickr groups through user studies and interviews [62], concluding that memory, narrative, relationships maintenance, self-representation, and self-expression are the five main motivations to join a group. Similarly, later work has come up with several alternative and partially overlapping classifications [37, 43].

Negoescu and colleagues have been among the main contributors in the study of Flickr groups. Initially, they manually categorized communities in *geographical*, *topical*, *visual*, and *catch-all* [40]. Following this initial categorization, they propose to detect hypergroups (i.e., groups of groups) based on the similarity of their topical focus, as determined by LDA [45]; on the opposite, Negi et al. have worked on splitting large Flickr communities in smaller subgroups using MoM-LDA on photo tags [39]. Negoescu et al. also analyzed groups in relation to their membership, with special attention to topicality and to peer-to-peer communication [41]. More recently they have discussed about how to represent Flickr groups according to the topics and tags defined by their members [42]. Supported by earlier studies on the same matter [62], they identify “real” groups as those motivated by self-expression and relationship maintenance, in contrast with those built around a specific topic (similarly to the socio-topical split we discuss later).

Motivated by a conceptual framework defined in earlier work [12], Cox et al. introduced the measure of "groupness" whose formulation takes into account the size of the group membership, the volume of photos, and the length of group description [16]. They propose to classify 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 any quantitative result. In partial contrast with previous work [42], their study suggests that small groups are more important than the big ones to improve social interaction dynamics because they operate at "human scale." The work was later extended [28] and the categorization was manually refined into four categories: *location-based*, *award*, *learning*, and *topical* groups.

Prieur et al. discuss the interplay between sociality and topicality in Flickr groups. By using PCA on a set of group features they detect the main components that characterize the group type. They find three main dimensions underlying as many types of groups: *social media-use*, *MySpace-like*, and *photo stockpiling* [47, 50, 51].

Social groups have also been described in terms the engagement of their members. From a quantitative perspective, the degree of involvement of the members in activities related to the group is varied and strongly dependent on group size [8]. Intra-group activity has been characterized in terms of item sharing practices [40], propensity of people to address other members' questions [66], or coherence of discussion topics [22]. Modeling inner activity of groups has helped in finding effective strategies to predict future group growth [31], recommend group affiliation, or improve the search experience on social platforms [45].

Groups have been studied also in other online platforms. User interaction patterns in groups extracted from YouTube, LiveJournal, DBLP, Orkut, and Yahoo Groups have been investigated in the past [7, 8, 38, 55]. In particular, the tendency to both topicality and sociality and the small-world nature of group interactions has been found in YouTube groups by Laine et al., who also envision in future work an analysis of the interplay between groups and the process of social influence [33].

Besides user-defined groups, the study of automatically detected groups through community detection algorithms has attracted much interest lately [54]. Detected communities are meant to represent meaningful aggregations of people where dense or intense social exchanges take place among their members [26]. Nevertheless, even if there is a variety of synthetic methods to verify the quality of detected communities [34], it is unclear whether such artificial groups capture any notion of community, as perceived by the users. If on the one hand the computation of cluster-goodness metrics over user-created groups can give useful hints about their structural cohesion [64], 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. Only recently researchers have been trying to address this question in a more systematic way [27, 29].

5.2.2 *Groups in Social Sciences*

Recent work in the field of computational social science tried to characterize communities according to the principles defined by well-known theories from social sciences. Activity and connectivity are heavily correlated with group size in several online social platforms [23, 26, 31], with a consistent patterns that recalls Dunbar's theory on the upper bound of around 150 stable social relationships for an average human [21]. Similarity between group members has been identified also as a factor driving the creation of social communities [59], also because of the tendency of social agents to aggregate according to the homophily principle [2]. However, similarity is not necessarily the strongest indicator for group stability or longevity, as diversity of content shared between group members is a major factor to keep the interest of members alive [36].

Social and thematic components of communities have been widely studied in social science, most of all within the common identity and common bond theory that will be discussed later in this chapter [49, 52, 53]. The principles behind the theory have never been translated into practical metrics to categorize groups, nor tested on large datasets, until recently. On the other hand, data-driven studies have investigated social and thematic components separately when characterizing groups [16]. Preliminary insights on the interplay between such dimensions have been given in exploratory work on Flickr, where signals of correlation between social density and tag dispersion in groups has been found [51] and where two different clusters emerge naturally when plotting the groups size against the number of internal links [9].

5.3 The Flickr Case-Study

As mentioned earlier, all the dimensions that we shall investigate in the following sections are quantitatively measured on a dataset extracted from Flickr. Its wide variety of user groups, the richness of interaction types, and the openness of the data make Flickr an ideal platform for this kind of study. Next, we shortly describe the main features of the dataset.

Users of Flickr can create, moderate and administer their own groups. Most groups allow users to join without an invite, whereas others are by invitation only and joining requires the administrator's permission. We consider a random sample of 500 K public groups created until the end of year 2008. For each of these groups, we extracted all the public information related to them (retrievable via the Flickr public API). All the data have been anonymized and processed in aggregate.

First, we collect the public information of group members about their social interactions:

- *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. Pictures can be stored in multiple pools, but only members of the group can upload a photo to its pool.
- *Subscriptions*. User u joins the group at a certain time.

Last, we collect photo *tags*. The primary set of photos from which we extract tags is the photo pool. In addition, 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. In the following, we will consider the three tag sets separately (pool, comments, favorites).

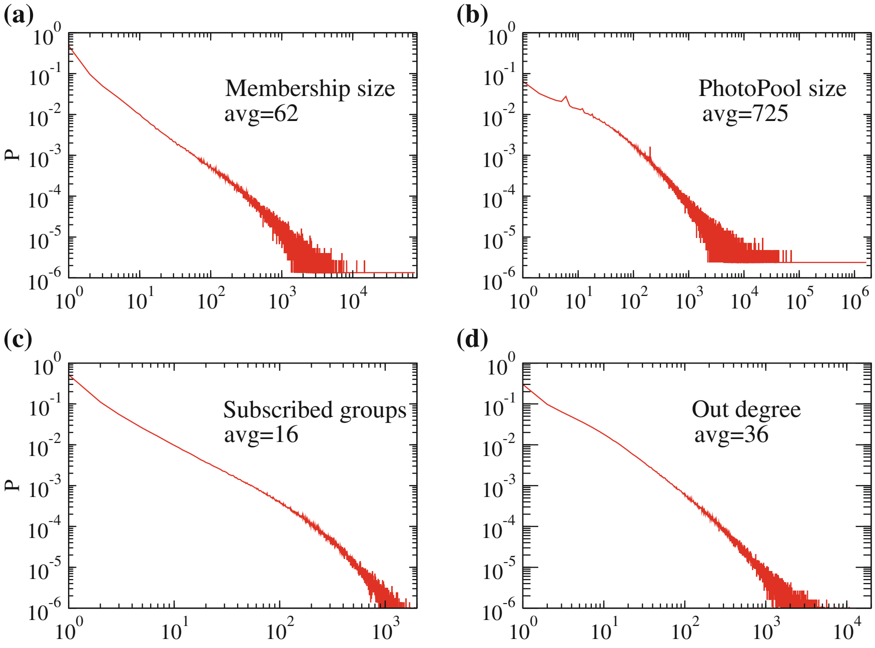

Fig. 5.1 Distributions (PDFs) of the characteristic dimensions of the dataset. Average values are reported in the plots. **a** Number of users in a group; **b** number of photos in a group pool; **c** number of groups a user is subscribed to; **d** out degree of the follower network induced by the users in our sample

The distributions of some of the main dimensions we consider in this study are reported in Fig. 5.1.

5.4 Space and Time Patterns of Groups

Actions within a group take place in a spatial and temporal context. Especially in online groups, where members can communicate even from long distance and maintain connections with relatively low cost, the spatial and temporal patterns can vary quite much. Next, we identify some metrics that can be used effectively to characterize communities along space (Sect. 5.4.1) and time (Sect. 5.4.2) [18], and we use them to classify groups in our Flickr dataset (Sect. 5.4.3).

5.4.1 Spatial Features

The first aspect we take into account is the location, namely the geographical position of the members or of the photos that are uploaded in the group pool. Geographical distribution of group members can indeed be correlated with the purpose of the group, being sometimes very localized (e.g., members of a photography club in the same city) and sometimes very broad (e.g., the club of Nikon camera owners). In his study on the geographical distribution of viewers of a given photo, Van Zwol [63] proposed three metrics to account for geographic sparsity. The first is the average over the geodesic distance geo_d between all the pairs of locations (Fig. 5.2a):

$$\begin{aligned} geo_d(lat_1, lon_2, lat_2, lon_2) \\ = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{lat_1 - lat_2}{2}\right) + \cos(lat_1) \cdot \cos(lat_2) \cdot \sin^2\left(\frac{lon_1, lon2}{2}\right)}\right), \end{aligned}$$

This metric scales quadratically with the number of points and it could be computationally prohibitive for large sets of locations. A way to overcome this issue is to estimate the dispersion by computing the standard deviation for the longitudes and latitudes separately and use them to build a bounding box around the centroid of the Cartesian coordinates (Fig. 5.2c). Then the Euclidean distance between the angles of the bounding box is considered as a measure of geographical dispersion. 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 map to very different distances. A direct solution to solve this problem is to use the geodesic distance instead (Fig. 5.2c). Still, even if the geodesic distance accounts for the curvature, it does not consider the Earth as a sphere, 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).

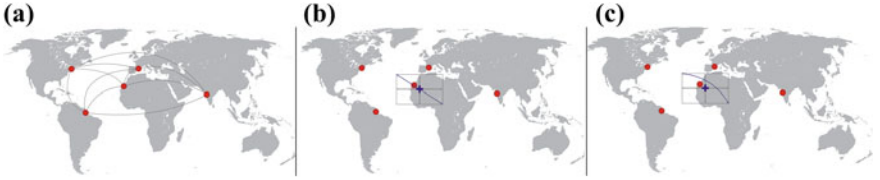

Fig. 5.2 Methods to measure dispersion of geolocated points (*red dots*) on a map. **a** Average of pairwise geodesic distances between points. **b** Diagonal of the bounding box defined by the standard deviations of latitude and longitude around the center of gravity (*blue cross*). **c** Same as (**b**) but considering the geodesic distance of the diagonal

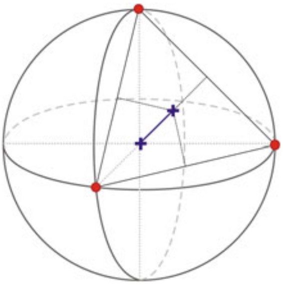

Fig. 5.3 Center-of-Earth distance method to measure dispersion of geolocated points (*red dots*) on the Globe. Points are translated into spatial Cartesian coordinates. The distance from their centroid to the center of the Earth (*blue segment*) is calculated as a measure of dispersion

To address these problems, we use the *Center-of-Earth distance* (coe_d) to directly measure geographical dispersion (Fig. 5.3). We consider each latitude-longitude pair as a polar-azimuth angle in the spherical coordinate system centered on the center of the Earth. We convert all the points into the three-dimensional Cartesian system. As all the points all lie on the spherical surface, their centroid always lies 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 so that its range is in [0, 1]. When just one point is available (or when multiple points overlap), the spread is maximally narrow ($coe_d = 1$), whereas points at the antipodes have a centroid residing exactly at the center of the Earth ($coe_d = 0$), yielding to maximum sparsity. Last, 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 addresses all the limitations of previous approaches because it has linear complexity, it takes into account the Earth's curvature and it considers the World as spherical.

5.4.2 Temporal Features

The temporal footprint of a group is represented by the sequence of events that happen within its boundaries. In the case of Flickr, for example, a photo upload or a new member joining the group could be events that compose the group's temporal trace.

Groups exhibit quite broad temporal patterns and the distributions of events in time are likely unique for each group instance. For this reason, high-level descriptors of the event timeline are needed to compare and cluster different groups according to their temporal profile. To do that, we rely on the statistical properties of the distribution of the events in time, specifically using four different properties: *central tendency*, *dispersion*, *skewness* and *burstiness*. In the following, we consider that all the events take place in a fixed, large time window $[0, T]$ that goes from the beginning of the system under study until the present time. Next, we define their meaning and propose metrics to capture each of them.

Central Tendency. In statistics, the central tendency or centrality of a distribution captures the tendency of the data to cluster around a central value. Given a sequence of timestamps in which group events occurred $(t_0 \cdots t_n) \in [0, T]$, with t_0 and t_n being the timestamps of the first and last events in the group, respectively, we define the central tendency as:

$$\mu_g = \frac{1}{n} \sum_{i=0}^{n} t_i. \tag{5.1}$$

This value expresses the central tendency of the event distribution in time and it is represented in the range [0, 1]. Values close to 0 indicate a high concentration of events at the beginning of the group lifetime, as opposed to a prevalence of events close to the present time for values approaching to 1.

Dispersion. A distinctive property of a distribution is its dispersion, namely how stretched or narrow a distribution is. To quantify this notion, we use a corrected version of the standard deviation that considers events on a normalized timeline:

$$\sigma_g = \sqrt{\frac{1}{n-1} \sum_{i=0}^{n} (t_i - \mu_g^t)^2 \frac{1}{n(1-\mu_g^t)\mu_g^t}}. \tag{5.2}$$

The range of values is [0, 1]. Groups with high central tendency have low dispersion, but groups with low dispersion could have also low central tendency. However, a non-corrected standard deviation would correlate heavily with the central tendency: a series of events with $\mu_g = 0.1$ can not have a dispersion higher than 0.5. To disentangle the two metrics, a 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.

Skewness. Skewness measures the asymmetry of the distribution with respect to its mean. It is calculated with the normalized difference between the median and the mean as follows:

$$\gamma_g = \frac{\mu_g - \text{median}_g}{\min(\mu_g, 1 - \mu_g)}. \tag{5.3}$$

Also in this case, the output 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 that are either smaller or larger than the median. The correction factor introduced in the denominator ensures the independence between the skewness and the central tendency, as we detail in the Appendix.

Burstiness. Last, we use a burstiness metric to measure the extent to which the group events happen simultaneously in big bursts. To capture this notion, we recur to the inter-event time ($\Delta_g^{t_{ij}} = t_j - t_i, i < j$). We refer to Δ_g^t as the overall series of inter-event times for a group g. The burstiness is defined as follows:

$$\Delta = \log_{10}(\mu(\Delta_g^t)) - \log_{10}(median(\Delta_g^t)). \tag{5.4}$$

The mean of all the inter-event times $\mu(\Delta_g^t)$ is equivalent to the total time between t_0 and t_n, divided by the number of events. The median of the inter-event times has values in the range [0, $\mu(\Delta_g^t)$]. Series with uniformly separated events have equal values of $\mu(\Delta_g^t)$ and median(Δ_g^t), whereas groups with a bursty behavior will have a median(Δ_g^t) that approaches 0.

5.4.3 Spatial and Temporal Groups in Flickr

Next, we apply the metrics of spatial and temporal characterization to the set of Flickr groups described in Sect. 5.3.

From the geographical perspective, we characterize groups using the single coe_d dispersion metric. However, the metric could be computed on different types of geolocated data: declared user location (in the user profile or in their IP address) and photo geotags. We do not consider the user geolocations for two reasons. First, some users do not provide their position in their own profile; additionally, the IP-based geolocation could be quite unreliable [63]. Last, our goal is to characterize groups with the information that is directly related to that group rather than to the users participating to them. For this reason, we consider the geotags attached to the photos uploaded to the group instead. As an example, consider a group that gathers tourists from all over the World who take pictures in Paris. In this case, we rather characterize the group as geographical narrow, as its focus is a single city, rather than describing the geographical dispersion of the member's locations.

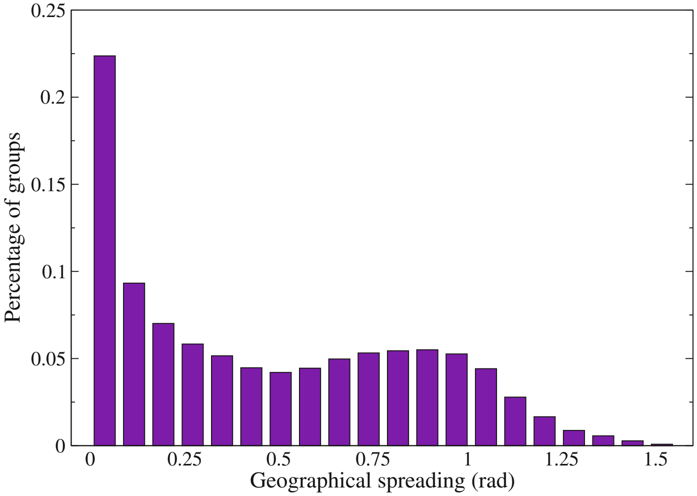

Fig. 5.4 Histogram of the ceo_d dispersion values of Flickr groups (values transformed in radians)

The application of the geographic dispersion metric with photo geotags yields the distribution over groups shown in Fig. 5.4. The histogram has a bi-modal distribution with local maximum around zero that includes the groups containing photos geographically near, and another local maximum around the 0.85 radians (≈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 functional to the purpose of the community.

To transition from a continuous value of coe_d to a discrete clustering of groups we apply the X-Means algorithm [46] over the monodimensional space of dispersion values, to avoid manual thresholding. X-Means is a variant of K-Means that allows for an automatic discovery of the optimal number of clusters K 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 metrics can be instantiated on two types of events, namely users joining the group and photos being uploaded in the group pool. Combining those two types of events with the four metrics we use to characterize the event distribution, we obtain eight distinct features. Similarly to the spatial clustering, we apply X-Means to this 8-dimensional feature space, obtaining three different clusters.

The average and standard deviation of every feature are shown in Table 5.1. The three features that are most discriminative are the dispersion and burstiness over users

Table 5.1 Average and standard deviation of every temporal feature for each of the clusters

	Photos				Users			
Clusters	Cent.	Disp.	Skew.	Burst.	Cent.	Disp.	Skew.	Burst.
Evergreen	0.42 ± 0.16	0.56 ± 0.14	0.49 ± 0.16	0.61 ± 0.21	0.47 ± 0.15	0.58 ± 0.13	0.48 ± 0.15	0.81 ± 0.15
Short-lived	0.03 ± 0.07	0.12 ± 0.16	0.60 ± 0.23	0.66 ± 0.21	0.05 ± 0.09	0.16 ± 0.16	0.58 ± 0.27	0.82 ± 0.13
Bursty	0.23 ± 0.16	0.56 ± 0.19	0.71 ± 0.21	0.57 ± 0.22	0.15 ± 0.11	0.60 ± 0.19	0.86 ± 0.15	0.62 ± 0.23

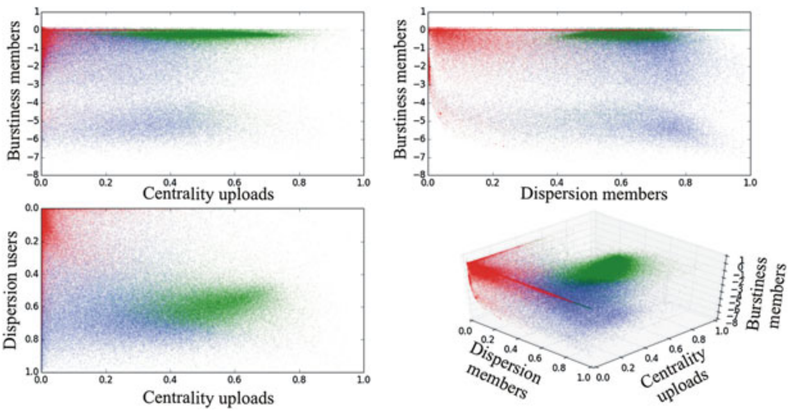

Fig. 5.5 Scatter plot of the groups with respect to the three most discriminative features for the clustering on the temporal dimensions. Bursty groups are depicted in *green*, evergreen in *blue*, and short lived in *red*

joining the group, and the centrality of the uploaded photos. A scatter plot of these three features for each cluster is reported in Fig. 5.5. After inspecting the clusters, we name them *evergreen*, *short-lived* and *bursty*. Next, we report their peculiar features.

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 experienced a low level 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 high centrality and by dispersion values around 0.5. They 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 communities of amateur and professional photographers interested in artistic portraits.

Bursty. The remaining 34 % of the groups belong to 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 and 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 regularly attract the attention of users.

The evolution of the number of users and photo uploads for the three most representative groups in each class is shown in Fig. 5.6.

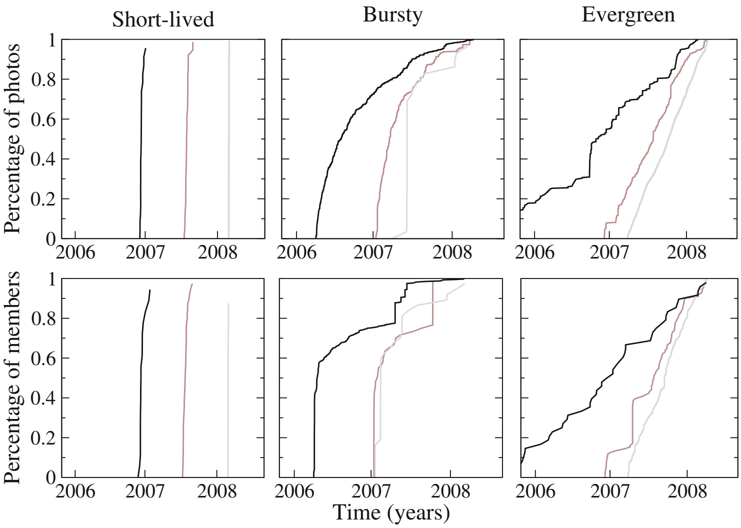

Fig. 5.6 Evolution of each of the three most representative groups in each temporal cluster. Values on the y-axis are normalized by the maximum values reached at the end of the time window of our dataset

5.5 Social and Topical Groups

5.5.1 Common Identity and Common Bond Theory

As mentioned in the introduction, well-established sociological studies have defined a connection between social groups and the process of formation of a social identity of the individual members [56]. The feeling of identity, or in other words the sense of belonging to a community, can be indeed very strong even in groups whose members do not know each other, as group identity can originate merely by defining a collection of people belonging to the same abstract category [60]. In extreme cases, the sense of identity can even emerge when members are randomly assigned to arbitrarily-defined communities [58]. Supporters of a political party, people who suffered from the same illness, members of a fan club, and people interested in the same hobby are all examples of groups that are defined by a common identity.

Groups that convey a strong identity are usually resistant to membership turnover, as individual members are interchangeable as long as the same sense of identity is preserved. However, this is clearly not the case for all the groups we can think of. For example, a person can join a group mainly because he has a direct friendship connection with a member, even without feeling a common identity with the group as a whole. As a result, if the latter leaves the group, the first is likely to quit as well [32, 52]. In this case, individual social links, more than an abstract notion of identity, constitute the backbone that allows the group to survive.

This duality of the groups' nature has been captured and discussed by Prentice in its formalization of the *common identity and common bond* theory [49] which states that, depending on the prevalent motivation of people to join, groups can be categorized as either *bond-based* or *identity-based*. Prentice assumes that the two types of groups have distinct and well-recognizable traits. 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 the members. Members whose participation is due to identity-based attachment may not directly engage with anyone and might even participate without revealing their identity. On the other hand, bond-based attachment is driven by personal relations with other members, and thus the main theme of the group may be disregarded. The two processes result in two different group types; for simplicity of exposition, in the following we will refer to those two categories as *social* and *topical* groups respectively.

In practice, groups can be formed from a mixture of bond- and identity-based attachment, even though very often they tend to lean on one aspect more. According to the theory, the group type is related with the *reciprocity* and the *topics* of discussion. Members of social groups tend to establish reciprocal interactions with other members, whereas interactions in topical groups are generally not directly reciprocated. Furthermore, topics of discussion in social groups tend to cover multiple subjects, while in topical groups discussions tend to be related to the group scope, covering specific topics only. 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. As we discussed, this makes social groups more vulnerable to turnover, since the departure of a person's friends may influence her own departure. On the opposite, topical groups are more open to newcomers and more robust to departures.

In recent years, the theory has been widely commented and elaborated by social scientists from a theoretical perspective and through small-scale experiments [52, 53, 61], but no rigorous methodology to distinguish the two types has been developed nor tested on large-scale datasets, until recently [27]. Next, we describe a technique to detect the group type based on the common identity and common bond theory. The method contributes to validate the theory itself but provides also a general framework for automatic classification of user groups in online social media.

5.5.2 From Theory to Metrics

It is possible to construct metrics to differentiate between the two types of groups by quantifying their reciprocity of interactions, and the topical width 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. We discuss how these metrics are combined in Sect. 5.5.4, with specific examples on our Flickr case-study.

Reciprocity. Reciprocity of interaction happens when a user sends any type of message to another user and, subsequently, the recipient responds with a new message. We define *intra-reciprocity* of a group g as:

$$r_g^{\text{int}} = \frac{E_g^{\text{int,rec}}/2}{E_g^{\text{int,rec}}/2 + E_g^{\text{int,nrec}}}, \tag{5.5}$$

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 $\left\langle r_g^{\text{int}} \right\rangle$ over all groups:

$$t_g = \frac{r_g^{\text{int}}}{\left\langle r_g^{\text{int}} \right\rangle}. \tag{5.6}$$

The larger the intra-reciprocity, the higher the probability that the group is social. 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:

$$u_g = \frac{r_g^{\text{int}} + 1}{r_g^{\text{ext}} + 1}. \tag{5.7}$$

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 towards people not belonging to the group.

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:

$$H(g) = - \sum_{t \in T(g)} p(t) \cdot \log_2 p(t), \tag{5.8}$$

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:

$$h_g = \frac{H(g)}{\langle H(f)\rangle_{|T(g)|=|T(f)|}}. \tag{5.9}$$

Activity. Even if, according to the common identity and common bond theory, activity is not a discriminative factor to discern social from 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:

$$a_g = \frac{E_g^{\text{int}}}{(D_g^{\text{in}} D_g^{\text{out}})/E}. \tag{5.10}$$

D_g^{in} and D_g^{out} are the total numbers of interactions originated by or targeted to members of the group g. E is the total number of interactions in the network. Values higher than 1 are obtained when 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 community is to compare density of its internal interactions with the density of interactions with the external world:

$$b_g = \frac{E_g^{\text{int}}/(s_g(s_g-1))}{E_g^{\text{ext}}/(2(N-s_g)s_g)}, \tag{5.11}$$

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 the density interactions between the group and the rest of the network.

5.5.3 Ground Truth of Social and Topical Flickr Groups

The socio-topical dimension we consider is a rather abstract concept; for this reason, a validation step is needed to check whether our metrics are able to correctly capture it. We resort to human editors to build a reliable ground truth of topical an social groups, under the assumption that the human capability of processing the semantics, aesthetics, and sentiment behind text and photos of a group allows for an easy discernment of social and topical groups. For the labeling, we randomly sampled groups that have (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 ensures that the selected groups are active above the expected values in a random case. After this selection we obtained over 34 K groups. The editors were asked to label groups after being presented with the following information:

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.

Comments. We provide the text of all the comments that are made between the group members. Comments are shown in chronological order and are grouped by thread, if they appear under the same photo. A link to the photo is also provided.

Tags. An alphabetically-sorted list of the 5 most frequent tags attached to the photos that group members commented on.

Editors 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 binary. Editors were asked to label groups based on well-defined guidelines extracted directly from the common identity and common bond theory [52]. The guidelines involve the inspection of two aspects. First, editors look at photos and comments based on the intuition that 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 [15], as opposed to topical groups, where the atmosphere is supposed to be more formal and impersonal [53]. Last, the editors inspect the photo tags and the group textual description to assess the semantic coherence that is typical of identity-based groups. Geo-referenced photos taken in a narrow geographical space can be an indication of high sociality, instead.

If both tags and comments are highly social or topical, then the label choice 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 they needed to get to a fairly clear decision.

Clarifying examples have been provided to the labelers to facilitate their task. For instance the "Airlines Austrian" group, tagged with "aircraft", "airport" and "spotting", that contains photos of airplanes from different countries in Europe is a clear example of a topical group. The "Camp Baby 2008" group, containing photos depicting people attending an event and interacting with each other with a friendly attitude is a social one; although the group has a specific topic and, as such, it contributes to the creation of the identity of its members, its social component is greatly predominant.

Multiple independent editors are asked to assess the quality of the extracted ground truth. A total of 101 groups were labeled by 3 people. The inter-labeler agreement, measured as Fleiss' Kappa, is 0.60, meaning that there exists good agreement between labelers. Once high agreement was assessed, we continued with individual labeling for a total of 565 distinct groups. We find the two types of groups being quite balanced in number, with around 48 % of social groups. One of the expectations is that bond-based groups should not be very large, as the human capacity for stable relationships is limited (as later discussed in Sect. 5.8). 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 172 members.

5.5.4 Group Type Prediction in Flickr

Before assessing experimentally the predictive power of the metrics, we inspect their properties to check how much their values differ between groups labeled as social or topical. In Fig. 5.7, we plot them as a function of the group size, to compare groups of similar sizes to draw unbiased conclusions.

We spot almost no differences in the number of photos (not shown), favorites, and contacts (as in Fig. 5.7b, c) between social and topical groups. The number of comments is, however, around 2 times higher in social groups than in topical groups of similar size (Fig. 5.7a). More differences are found when looking at relative activity (Fig. 5.7d–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 up to over 10 times higher than for topical groups.

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 (Fig. 5.7j–o), in line with the theory. There is no difference in reciprocity of contacts, plausibly because contacts do not strongly reflect personal relations between connected users. Possibly, since contacts do not need to be reciprocal, users often "follow" people they do not know and do not actively interact with. Finally, we observe much higher values of entropy and normalized entropy in social groups than in topical ones (Fig. 5.7p, q, s, t). This holds for the tags extracted

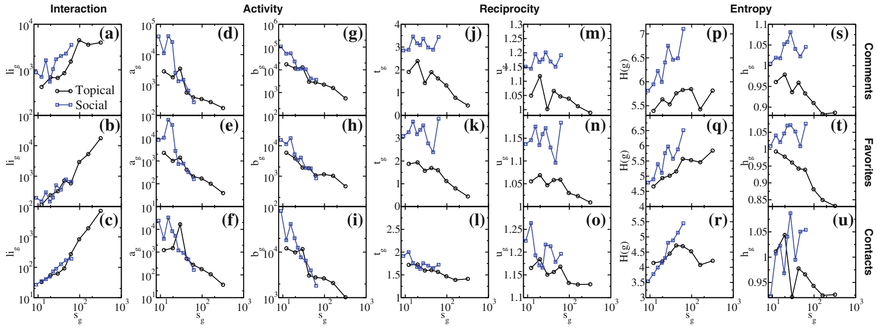

Fig. 5.7 Average (**a–c**) number of interactions, (**d–i**) activity (a_g, b_g), (**j–o**) reciprocity (t_g, u_g), and (**p–u**) entropy ($H(g)$, h_g) of topical (*black circles*) and social (*blue squares*) groups as a function of their size. Each point corresponds to 30 groups. Commenting, favoriting and social linking (contacts) are the three interaction types considered

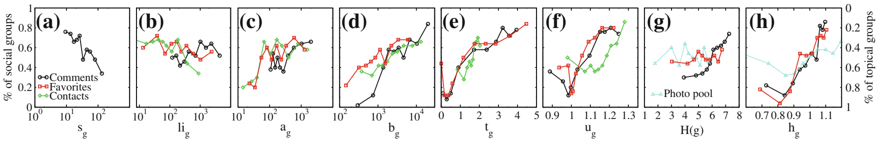

Fig. 5.8 Dependence of fraction f of groups labeled as social on various metrics: computed considering comments, favorites, contacts and photo pools. Each point corresponds to 50 groups

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 engage in interactions covering many different topics, whereas members of identity-based groups focus their interactions on specific topics. Apparently though, this does not hold for the tags extracted from photo pool of the group (Fig. 5.7r, u). This might be explained by the fact that the content of the photo pool does not always reflect well the interactions and relations between members of the group.

We also look at the fraction of groups labeled as social with respect to their size, activity, reciprocity, and entropy (Fig. 5.8). The group size correlates negatively, as expected (Fig. 5.8a). The correlations with the number of interactions and relative activity a_g are quite weak (Fig. 5.8b, c), whereas surprisingly there is a strong correlation on relative activity b_g (Fig. 5.8d). 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. That might relate to the difficulty of joining bond-based groups due to strong relations existing between their members and to the high investment that is required to create such relations with them [52]. Direct reciprocity of interactions, with the exception of contacts, correlates strongly with social groups (Fig. 5.8e, f). Furthermore, we find that the entropy of tags correlates with group sociality, but entropy based on other sources does not (Fig. 5.8g). However, the normalized entropy performs better and correlates strongly when computed on tags extracted from both comments and favorites (Fig. 5.8h).

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 our ground truth. The easiest approach to use is a linear combination. To do that, we select the features that directly implement the sociological theory: t_g, u_g, and h_g. Each of them is computed for the 3 different interaction types and bags of tags, yielding a total of 9 values. We transform the values into their t-statistics by subtracting the average and dividing them by the standard deviation. We weight the normalized scores evenly and we sum them up to obtain a single *sociality score* S_g. All of the components are supposed to score high for social groups, therefore the higher the value of the final 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 is selected.

The second approach relies on machine-learning methods trained with the features we have identified, using the labeled groups in ground truth as training examples. The classifier 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 other popular classification approaches. For this supervised approach we use a wider set of features than the one we used in the linear combination, namely: s_g, 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.

The ratio of social groups increases quickly with the score S_g, as illustrated in Fig. 5.9a, suggesting that the features embedded in the score are able to capture the nature of the groups to some extent. When the score is around zero, groups can be characterized by a mix of social or topical aspects, and a decision on the predominant nature of the group is more difficult. If we fix the threshold for the S_g value 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 Fig. 5.9a. 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 (Fig. 5.9b). 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. Consistently, this drop occurs also in the pairwise agreement between human labelers, measured as a ratio of groups that have been given the same label. Apparently, this is a shortcoming of the binary classification coming from the common identity and common bond theory itself, rather than of the features or the prediction framework.

We compare the performance of the two approaches through ROC curves (Fig. 5.9c), 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 5.2.

Finally, to shed light on which are the most predictive features, we rank them using Chi-square feature selection. The top 5 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 5.2. The top 4 most predictive features correspond directly to the expectations of the theory. Reciprocity-based metrics and normalized

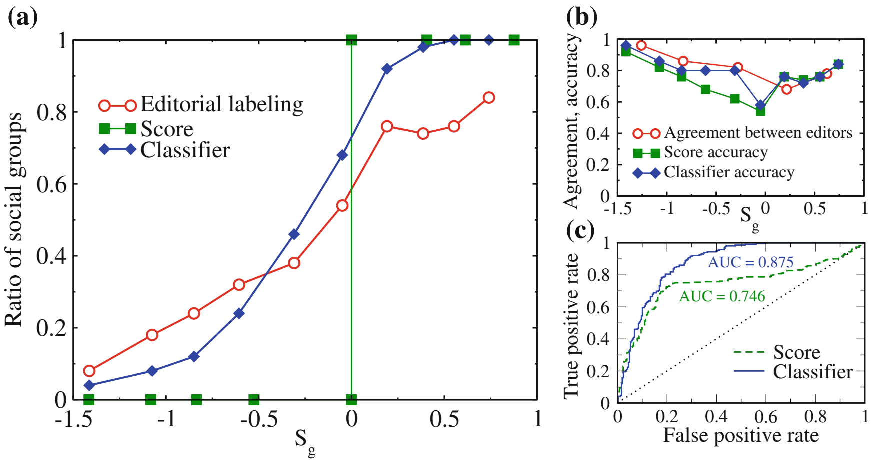

Fig. 5.9 Prediction of group type (social vs. topical). **a** Ratio of groups classified as social (by the labelers, by the linear combination methos with threshold at 0, and by the classifier) versus the sociality score S_g. **b** Accuracy of the two prediction techniques and agreement between labelers against the S_g values. **c** ROC curves for the prediction with the two different techniques

Table 5.2 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
Classifierχ^2_{top5}	0.803	0.872

entropy are significantly more predictive than other features. The high position of relative activity b_g^{com} is instead more unexpected.

5.6 Towards a Comprehensive View on Group Types

We have laid down the foundations for a group characterization along the spatial, temporal, and socio-topical aspects separately. A natural question that arises is whether there are some cross-dimension relationships between group types, or in other words, if different clusters of groups in one dimension correspond predominantly to some other type of group in the other dimension. Blending all the metrics in a single model could 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.

Table 5.3 Percentage of groups in each intersection between clusters

	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

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. Two spatial (geo-wide and geo-narrow), three temporal (evergreen, short-lived, and bursty), and two socio-topical (social and topical) classes yield 12 possible combinations of classes. The relative volume of the Flickr groups in our sample for each of them is reported in Table 5.3. Interesting 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 group right after its creation and the group would probably experience a activity bursts in correspondence to the real-life 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. Last, 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; in other words, geo-width is an indicator of a better chance of the group to survive longer.

5.7 Declared Versus Detected Groups

Community detection techniques have been largely employed in recent years to describe the structure of complex social systems [54]. The need for a clearer assessment of the *meaning* of the detected clusters has been often expressed from different angles [34, 64], but never completely satisfied by empirical analysis. Here we contribute to shed light on this matter by comparing user-generated groups (*declared* groups) with groups detected algorithmically (*detected* groups).

To automatically find communities, we apply the OSLOM community detection algorithm [35] 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 community detection benchmarks [34] and it outperformed other algorithms we tested. OSLOM detected 646 K groups, overall.

First, we check the tendency of detected communities towards sociality or topicality with another round of manual annotation. Three independent annotators labeled

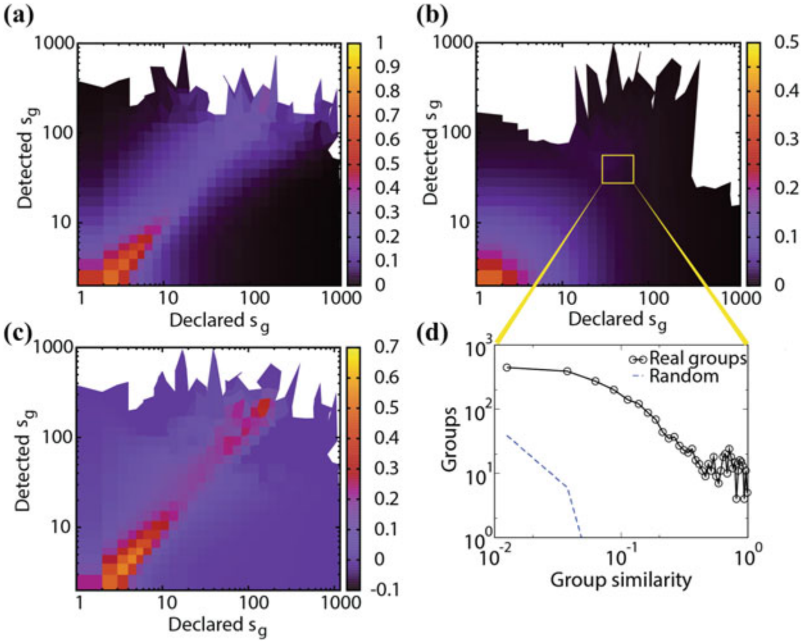

Fig. 5.10 Overlap between declared and detected groups. Jaccard similarity between the member sets of declared and detected groups as a function of their sizes in **a** the actual data, **b** in a null model with randomized groups, and **c** difference between the two. **d** Histogram of the similarity values for a sample of groups in the diagonal

126 distinct detected groups obtaining a Kappa value for detected groups around 0.44. The lower agreement than the one reached for the declared groups is partially determined by the lack of information about the group's profile, not available for detected groups. Among detected groups almost 69 % are labeled as social (vs. 48 % in declared groups).

We then compare the size of declared and detected groups. The size distribution is heavy-tailed and close to power-laws in both cases (not shown) but 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 a function of the size of groups in Fig. 5.10a. For the purpose of comparison with a null model, in Fig. 5.10b 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 Fig. 5.10c, meaning that indeed detected groups are, to some extent, similar to the declared ones. Further insights are given by the distribution of similarities of pairs of groups extracted from a small sector of the diagonal, having between 32 and 64 members (Fig. 5.10d). Unlike in the randomized scenario, there are multiple detected groups which overlap significantly with declared groups, and that randomized groups do not show this pattern.

We can therefore conclude that, in some cases the community detection algorithm finds groups which are very similar to the ones defined by the users. Nevertheless, substantial overlap is found for just a small percentage of groups and most of the group pairs have similarity close to 0. The average similarity of detected groups to the best-matching declared groups is 0.082, while for the randomized detected groups is 0.058, only slightly lower.

Additionally, we picked 50 detected groups among the ones that are the most similar to declared groups. 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 that happens is not trivial.

5.8 The Barrier of Membership Size

The membership size is another important feature of groups, as group size necessarily affects the dynamics of interaction between members.

This intuitive concept has been discussed in depth by the anthropologist Robin Dunbar. In a study he performed in 1992, he measured the correlation between the neocortical volume of primates and the typical size of their social communities [20]. The limits in the size of social groups of primates has been explained primarily by the limited amount of time that an individual could dedicate to social grooming in addition to the organizational issues that arise when communities grow in size and that can be tackled only by enforcing norms that help to maintain them stable.

By extrapolating from the result obtained on primates, Dunbar theorized that the limit of community size for human beings should lie roughly in a ballpark of 100–250, being larger groups too demanding to manage in terms of cognitive efforts for an average person. The anecdotal figure, often presented as the *Dunbar number* is that the maximum size of groups that an individual can manage with reasonable cognitive effort is 150.

The advent of online social media has provided large dataset to verify this theory at scale. One of the most notable attempts has been done by Goncalves et al. [23] on the Twitter by measuring the average social strength ω_i^{out} of each individual i on the mention network:

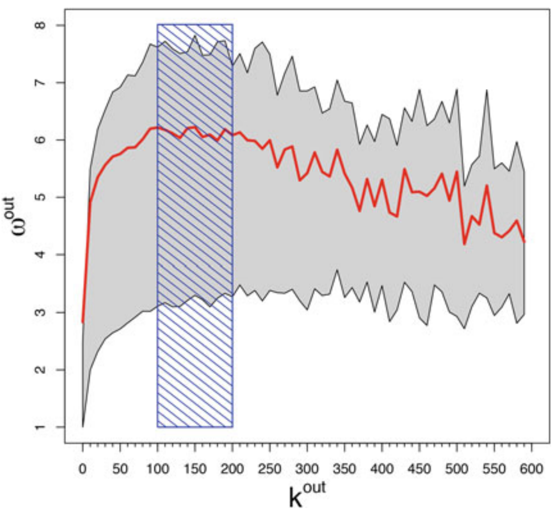

Fig. 5.11 Out-weight ω_{out} as a function of the out-degree in a Twitter mention network. The *red line* corresponds to the average out-weight, while the *gray shaded area* illustrates the 50 % confidence interval. Figure and caption taken from the original publication, courtesy of the authors [23]

$$\omega_i^{out}(T) = \frac{\sum_i w_{ij}(T)}{k_i^{out}}, \tag{5.12}$$

where w_{ij} is the weight of the edge between users i and j, the weight representing the number of messages exchanged within a time window T, and k_i^{out} is the outdegree of user i, namely the overall number of people he has mentioned durint that time window. In short, ω^{out} represents the amount of attention that the individual pays to her social partners in a certain time frame. Averaging the value of ω^{out} for all the users with the same value of k_{out} and plotting the resulting values against k_{out} results in the trend displayed in Fig. 5.11. The average strength gradually increases until it reaches its maximum between 100 and 200 contacts, signaling that a maximum level of social activity has been reached. Beyond that point, an increase in the number of contacts can no longer be sustained with the same amount of dedication, as Dunbar theorized.

The strong evidence that supports Dunbar's theory in the Twitter scenario by looking at egocentric networks, can be also corroborated with a group-centered perspective. If Dunbar's hypothesis holds, groups that are larger than a certain size will have much lower interaction density between their members than smaller groups. To capture that, we use the activity measure a_g that we have presented in Sect. 5.5.2 and that we report again for the reader's convenience:

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

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, and E is the total number of

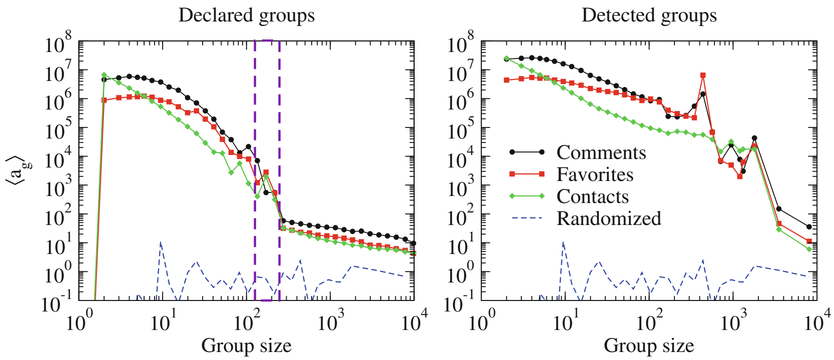

Fig. 5.12 Group activity a_g as a function of the group size for Flickr groups. Activity is measured considering three types of dyadic interactions that might happen inside the group: commenting, favoriting, and creation of following contacts. For the purpose of comparison, we also recompute a_g in randomized null model where the members of groups are reshuffled but their size preserved

interactions in the network. The overall value is higher than 1 when the number of intra-group interactions is higher than the same number expected in a random scenario. When averaging a_g among Flickr groups with the same size, similarly to the analysis that has been conducted on Twitter, we obtain the curve shown in Fig. 5.12.

We observe that the activity decays almost monotonically by construction of the metric: the larger the group, the higher the likelihood of the density of its internal interactions to drop closer to values expected in a random case. However, when measuring a_g for declared groups, a sharp drop of activity occurs for groups with size between 100 and 250. This clearly means that after a certain size, the density of activity within the group cannot be sustained given the high number of participants. When running the same experiment on detected groups, the activity drop is steady and much more moderate. That happens because community detection algorithms tend by design to output node clusters with high numbers of connections between them.

Individuals with thousands of online social contacts are frequent in online social networks; likewise, groups with a membership size beyond the Dunbar number exist as well and indeed there is a large number of groups with thousands of members in Flickr. Those groups however tend either to be pure manifestations of social identity, representing more “social labels” than actual social aggregations (e.g., the group of Canon camera owners), or they are necessarily fragmented in smaller, more active communities. For this reason, when characterizing a group, size is yet another important feature to take into account to reach an unbiased understanding of the group’s nature.

5.9 The Role of Groups in Other Social Phenomena

Similarly to social links, groups are structural constituent of the social fabric that mediates most of the interaction dynamics between people. For this reason, the structure of the social network and the phenomena that occur over it are deeply intertwined. Nevertheless, studies in the areas of graph mining and social network analysis are too often conducted in separate sub-branches. One example is the relationship between the study of communities and research on information diffusion. As cleverly noted by Easley and Kleinberg [17], 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, communities usually aggregate people that share some common trait and therefore are more similar to each other than to the rest of the network. The more a community is dissimilar to the external world, the higher the probability of a piece of information that generates inside it to never cross the group borders. In other words: *"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"* [17].

Recently, this idea inspired the work of Barbieri et al. [10] who leveraged data on information cascades to detect hidden communities. Given a directed social graph and a set of information cascades observed over it, they propose a stochastic mixture membership generative model to detect communities of nodes that can *explain* such cascades.

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 easily. But what if a social (instead of topical) community is reached by the information cascade? To shed light on this matter we have run an experiment to check information cascades in relation to the types of groups we identified earlier [18]. To do that, we rely on a well-established work by Cha et al. [13, 14] that uses Flickr to analyze information propagation. They define the process of information diffusion using the favorite information. A piece of information propagates from user u_1 to user u_2 when all the following conditions 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 explore 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 to a group pool has a diffusion that is predominantly restricted to that group or spreads beyond the group boundaries. 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.

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 these 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 can be safely disregarded in this type of study.

Given this experimental setup, we compute a pair of values for each cascade. Consider $A_{g,p}$ 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 set of members of group g. We define:

$$c_{g,p} = \frac{|A_{g,p} \cap M_g|}{|M_g|} \tag{5.13}$$

$$s_{g,p} = 1 - \frac{|A_{g,p} \cap M_g|}{|A_{g,p}|} \tag{5.14}$$

The *coverage* $c_{g,p}$ measures how much the group is covered by the information cascade, the portion of group membership that is affected by the spreading process. The *external spreading* $s_{g,p}$ measure, instead, is designed to capture how much more the information spreads outside the group. An example of a cascade is given in Fig. 5.13.

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 different group types, we consider the values for each of the group types separately and we compute the average values at fixed group size, to account for any effect possibly given by group dimensionality. The results are shown in Fig. 5.14.

On the socio-topical axis, the difference between different types of group is slight but noticeable, with the topical groups having more coverage and less external spreading (except for a small range of group sizes). This supports the intuition reported in

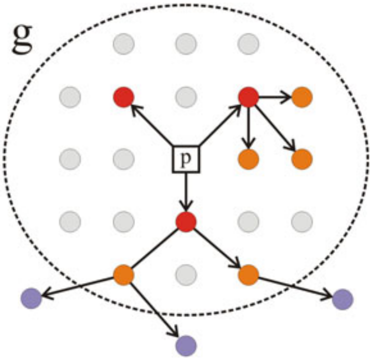

Fig. 5.13 Example of diffusion tree for a photo p uploaded into the photo pool of group g. *Circles* represent users and the *dashed line* marks the boundaries of the group. *Red circles* are the root users. In this example 8 users out of 20 members are nodes of the diffusion tree, leading to $coverage_{g,p} = 0.4$. 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 $externalCoverage_{g,p} = 0.27$

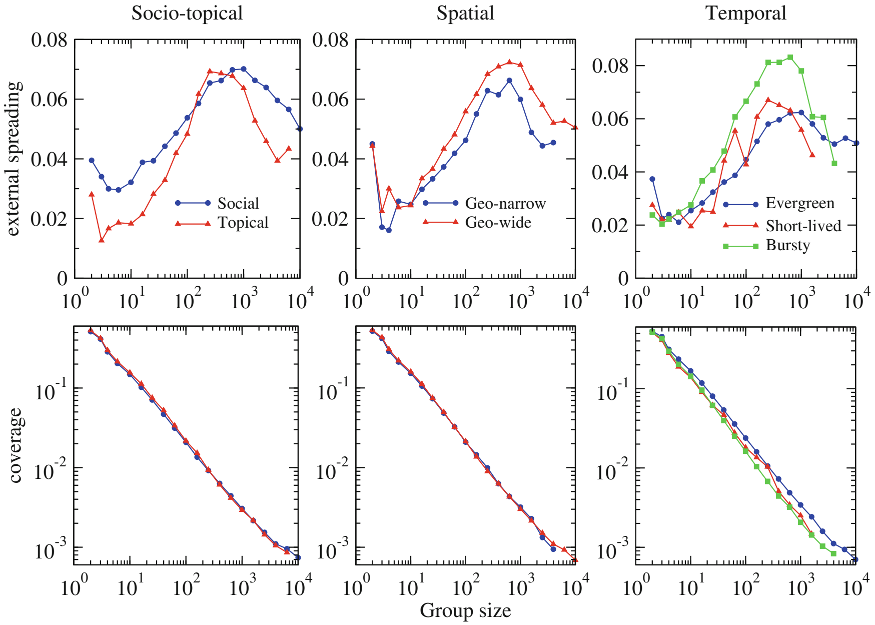

Fig. 5.14 Information diffusion in different group types. Average values of coverage and external spreading for all groups with same number of members. Groups are analyzed according to the three dimensions separately (socio-topical on the *left*, spatial in the *middle*, and temporal on the *right*)

previous work that identifies the boundaries of topical groups as harder to cross by information cascades. This is somehow expected when members of topical groups share interests which are narrow enough to be limited predominantly to the groups members. Conversely, 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 for the time dimension. On average, the evergreen groups have more coverage than the short-lived or the bursty, 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.

5.10 Are Groups the Missing Link Between Atomic Interactions and Emergent Social Phenomena?

Social networks fall into the category of complex systems, where relationship between atomic components give rise to an emergent behaviour that cannot be inferred or modeled directly from the composition of the individual parts. Complex processes in networks have been studied in several fields including physics, biology, and computer science. Also social scientists have been discussing the so-called *micro-macro* problem for long time. That refers to the duality (and often the incoherence) between the behaviour of an individual actor or of its interpersonal dyadic relations and the behaviour of the masses. In his book [11] Peter Blau, commented on this challenge:

> The problem is to derive the social processes that govern the complex structures of communities and societies from the simpler processes that pervade the daily intercourse among individuals and their interpersonal relations.

Later, in an updated introduction to the same book, he states:

> I thought that this microsociological theory could serve as a foundation for building a macrosociological theory; I no longer think this is true. The reason is that microsociological and macrosociological theories require different approaches and conceptual schemes, and their distinct perspective enrich each other.

Groups fall exactly in between the micro and macro scales, being manifestation of a collective identity that emerges from a limited number of individual motivations. The important role of groups in bridging different scales motivates even more the need for a nuanced characterization of their multiple facets.

We have contributed to fill this gap 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. We identify two main classes of spatially-characterized groups (geo-narrow and geo-wide) and discover three major patterns of their temporal activity (evergreen, bursty, and short-lived). By transposing the concepts of the common identity and common theory into metrics of reciprocity, activity, and topical diversity we are able to accurately tell apart social from topical groups. 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 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. We have also discussed the structural effect that group size in shaping the amount of activity within members, thus giving to groups of different sizes different relevance to the aspect of the construction of a social identity versus being vehicles of social bond construction. 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 crossing the boundaries of groups more than topical and evergreen groups, that instead tend to retain more information within them.

Besides carrying on detailed studies about all the facets of groups' structure and dynamics, it is equally important, as Blau wisely suggests, to corroborate and complement the findings of studies focusing at the group level by doing research on related social structures or dynamics. Especially, the socio-topical dichotomy coming from the principles of the common identity and common bond theory has been spotted also at the level of social link, without the need of fixing any apriori classes. In our recent work [4] we focus on dyadic conversations in Flickr (represented by mutual commenting on photos), trying to interpret individual conversational exchanges under the light of Blau's social exchange theory [11], stating that every dyad is a repeated set of exchanges of different types of *non-material resources* such as knowledge, social support or manifestation of approval. To associate each message to those non-material resources, we developed a method that combines topic detection with the analysis of reciprocation in conversations, motivated by the assumption that conversations might touch upon several topics but tend to exchange the same type of resource all along. This assumption has been derived as a theoretical necessity in the exchange of status [24], has been shown to exist in the case of social support [5]. The interesting aspect of the method is that, differently from classic classification approaches, the number of resources is not specified in input, allowing the discovery of the main non-material resources exchanged in any conversation network.

The application of the method on the Flickr conversation network finds two well-distinct domains, namely the ones of *status exchange* and *social support*, the first being associated to expression of appreciation or esteem for each other's work (e.g., "very nice shot, you are a good photographer!") and the second one representing everyday minute exchange or chit chat with some emotional evaluation (e.g., "how is your dad? I hope he is feeling better now"). The parallel between the socio-topical partition of groups is striking and, although the outcomes of the two different methods have not been directly compared yet in a quantitative way, it is surprising to get concordant results from an unsupervised method focused on atomic dyadic interactions and a supervised classification of groups.

Future work aimed at understanding social structures, either online or offline, should tap right into this direction: different social phenomena such as formation of groups and diffusion of information should be studied no longer in separation, as they are manifestations of the same complex entity. In this setting, groups might represent a key tile to bridge between the micro and macro scales of social interactions.

Acknowledgments We thank Przemyslaw Grabowicz, David Martin-Borregón, Rossano Schifanella, Bogdan State, Alejandro Jaimes, and Ricardo Baeza-Yates for the research work they have conducted jointly with the author and that is summarized in this chapter. This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975.

Appendix

Correction Parameter for Standard Deviation

Standard formulation of standard deviation is:

$$\sigma^2 = \sqrt{\frac{1}{N-1}\sum(t-\mu)^2} \tag{5.15}$$

Given a list N values t that can assume 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:

$$\sum(t-\mu)^2 = N \cdot p \cdot (0-\mu)^2 + N \cdot q \cdot (1-\mu)^2 \tag{5.16}$$

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

$$\begin{aligned}\sum(t-\mu)^2 &= N \cdot (1-\mu) \cdot (0-\mu)^2 + N \cdot \mu \cdot (1-\mu)^2 && (5.17)\\ &= N \cdot (1-\mu)\mu^2 + N \cdot \mu(1+\mu^2-2\mu) && (5.18)\\ &= N\mu(1-\mu) && (5.19)\end{aligned}$$

Therefore, being $N\mu(1-\mu)$ the maximum value for $\sum(t-\mu)^2$, we use it as normalization factor in Formula 5.2.

Correction Parameter for Skewness

Under a Bernoulli distribution with 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:

$$median = \begin{cases} 0 & if \quad p(0) > p(1) \\ 0.5 & if \quad p(0) = p(1) \\ 1 & if \quad p(0) < p(1) \end{cases} \tag{5.20}$$

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_{diff}) given the mean μ as follows:

$$maxdiff = \begin{cases} 1-\mu & if \quad p(0) < p(1) \\ \mu & if \quad p(0) > p(1) \end{cases} \tag{5.21}$$

Under a Bernoulli distribution taking values 0 and 1, the mean is equal to $p(1)$. Also, $p(0)$ is equal to the remaining $1-\mu$. Given that, we can rewrite the equation as:

$$max_{diff} = \begin{cases} 1-\mu & if \quad 1-\mu < \mu \\ \mu & if \quad 1-\mu > \mu \end{cases} \tag{5.22}$$

that can be finally rewritten as:

$$max_{diff} = \min(1-\mu, \mu) \tag{5.23}$$

which we use it as normalization factor in Formula 5.3.

References

1. Aiello LM, Barrat A, Cattuto C, Schifanella R, Ruffo G (2012) Link creation and information spreading over social and communication ties in an interest-based online social network. EPJ Data Sci 1(12):1–31
2. 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):9:1–9:33
3. Aiello LM, Deplano M, Schifanella R, Ruffo G (2012) People are strange when you're a stranger: impact and influence of bots on social networks. In: Proceedings of the 6th AAAI international conference on weblogs and social media, ICWSM'12. AAAI, pp 10–17

4. Aiello LM, Schifanella R, State B (2014) Reading the source code of social ties. In: Proceedings of the 2014 ACM conference on web science, WebSci'14. ACM, New York, pp 10–17
5. Antonucci T, Fuhrer R, Jackson J (1990) Social support and reciprocity: a cross-ethnic and cross-national perspective. J Soc Pers Relatsh, 7(4):519–530
6. 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 Press, New York, p 44
7. 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
8. Backstrom L, Kumar R, Marlow C, Novak J, Tomkins A (2008) Preferential behavior in online groups. In: Proceedings of the international conference on web search and web data mining—WSDM'08. ACM, New York, pp 117–128
9. 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, number 08391 in Dagstuhl seminar proceedings, Dagstuhl, Germany. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany
10. 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
11. Blau PM (1964) Exchange and power in social life. Transaction Publishers, New Jersey
12. Butler B (1999) When a group is not a group: an empirical examination of metaphors for online social structure. Ph.D. thesis, Carnegie Mellon University, Pittsburgh
13. 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
14. 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
15. Collins NL, Miller LC (1994) Self-disclosure and liking: a meta-analytic review. Psychol Bull 166(3):457–475
16. Cox A, Clough P, Siersdorfer S (2011) Developing metrics to characterize Flickr groups. J Am Soc Inf Sci Technol 62:493–506
17. David E, Jon K (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, New York
18. David M-B, Aiello LM, Grabowicz P, Jaimes A, Baeza-Yates R (2014) Characterization of online groups along space, time, and social dimensions. EPJ Data Sci 3(1):8
19. 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
20. Dunbar RIM (1992) Neocortex size as a constraint on group size in primates. J Hum Evol 22(6):469–493
21. Dunbar RIM (1998) The social brain hypothesis. Evol Anthropol 6:178–190
22. 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
23. Goncalves B, Perra N, Vespignani A (2011) Modeling users' activity on twitter networks: validation of Dunbar's number. PLoS ONE 6(8):e22656, 08
24. Gould RV (2002) The origins of status hierarchies: a formal theory and empirical test. Am J Sociol 107(5)
25. Grabowicz PA, Eguíluz VM (2012) Heterogeneity shapes groups growth in social online communities. Europhys Lett 97(2):28002

26. 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
27. 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
28. Holmes P, Cox AM (2011) Every group carries the flavour of the admins, leadership on Flickr. Int J Web Based Commun 7(3):376–391
29. Hric D, Darst RK, Fortunato S (2014) Community detection in networks: structural clusters versus ground truth. arXiv:1406.0146
30. Kairam S, Brzozowski M, Huffaker D, Chi E (2012) Talking in circles: selective sharing in Google+. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI'12. ACM, New York, pp 1065–1074
31. 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
32. Krackhardt D, Porter LW (1986) The snowball effect: turnover embedded in communication networks. J Appl Psychol 71(1):50–55
33. 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), January 2011, pp 1–10
34. Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78:046110
35. Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PLoS One 6(4):e18961, 04
36. 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
37. 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
38. 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
39. 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
40. 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
41. 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
42. Negoescu RA, Gatica-Perez D (2010) Modeling Flickr communities through probabilistic topic-based analysis. Trans Multi 12(5):399–416
43. 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
44. Park N, Kee KF, Valenzuela S (2009) Being immersed in social networking environment: Facebook groups, uses and gratifications, and social outcomes. Cyberpsy Behav Soc Netw 12(6):729–733
45. Negoescu RA, 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
46. 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 Publishers Inc, San Francisco, pp 727–734

47. Pissard N, Prieur C (2007) Thematic vs. social networks in web 2.0 communities: a case study on Flickr groups. In: Algotel conference
48. Porter CE (2004) A typology of virtual communities: a multi-disciplinary foundation for future research. J Comput Med Commun 10(1)
49. Prentice DA, Miller DT, Lightdale JR (1994) Asymmetries in attachments to groups and to their members: distinguishing between common-identity and common-bond groups. Personal Soc Psychol Bull 20(5):484–493
50. Prieur C, Cardon D, Beuscart J-S, Pissard N, Pons P (2008) The strength of weak cooperation: a case study on Flickr. CoRR, arXiv:0802.2317
51. Prieur C, Pissard N, Beuscart JS, Cardon D (2008) Thematic and social indicators for Flickr groups. In: Proceedings of ICWSM
52. Ren Y, Kraut R, Kiesler S (2007) Applying common identity and bond theory to design of online communities. Organ Stud 28(3):377–408
53. Kai S (2002) Common bond and common identity groups on the internet: attachment and normative behavior in on-topic and off-topic chats. Gr Dyn Theory Res Pract 6(1):27–37
54. Santo F (2010) Community detection in graphs. Phys Rep 486(3–5):75–174
55. Spertus E, Sahami M, Buyukkokten O (2005) Evaluating similarity measures: a large-scale study in the Orkut social network. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, KDD'05. ACM, New York, pp 678–684
56. Tajfel H (1981) Human groups and social categories. Cambridge University Press, Cambridge
57. Tajfel H (1982) Social identity and intergroup relations. Cambridge University Press, Cambridge
58. Tajfel H, Billig MG, Bundy RP, Flament C (1971) Social categorization and intergroup behaviour. Eur J Soc Psychol 1:149–178
59. Tang L, Wang X, Liu H (2011) Group profiling for understanding social structures. ACM Trans Intell Syst Technol 3(1):15:1–15:25
60. Turner JC (1985) Social categorization and the self concept: a social cognitive theory of group behavior. In: Lawler EJ (ed) Advances in group process. JAI, pp 77–122
61. Utz S, Sassenberg K (2002) Distributive justice in common-bond and common-identity groups. Gr Process Intergr Relat 5(2):151–162
62. 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
63. Van Zwol R (2007) Flickr: who is looking? In: IEEE/WIC/ACM international conference on web intelligence, WI'07. IEEE Computer Society, pp 184–190
64. Yang J, Leskovec J (2012) Defining and evaluating network communities based on ground-truth. CoRR, arXiv:1205.6233
65. 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
66. Welser HT, Gleave E, Fisher D, Smith M (2007) Visualizing the signatures of social roles in online discussion groups. J Soc Struct 8(2)