Distinguishing Topical and Social Groups Based on Common Identity and Bond Theory

Przemyslaw A. Grabowicz[§][*]   Luca Maria Aiello[†]   Víctor M. Eguíluz[§]   Alejandro Jaimes[†]

[§]{pms,victor}@ifisc.uib-csic.es
IFISC (CSIC-UIB)
Palma de Mallorca, Spain

[†]{alucca, ajaimes}@yahoo-inc.com
Yahoo! Research
Barcelona, Spain

ABSTRACT

Social groups play a crucial role in social media platforms because they form the basis for user participation and engagement. Groups are created explicitly by members of the community, but also form organically as members interact. Due to their importance, they have been studied widely (e.g., community detection, evolution, activity, etc.). One of the key questions for understanding how such groups evolve is whether there are different types of groups and how they differ. In Sociology, theories have been proposed to help explain how such groups form. In particular, the common identity and common bond theory states that people join groups based on identity (i.e., interest in the topics discussed) or bond attachment (i.e., social relationships). The theory has been applied qualitatively to small groups to classify them as either topical or social. We use the identity and bond theory to define a set of features to classify groups into those two categories. Using a dataset from Flickr, we extract user-defined groups and automatically-detected groups, obtained from a community detection algorithm. We discuss the process of manual labeling of groups into social or topical and present results of predicting the group label based on the defined features. We directly validate the predictions of the theory showing that the metrics are able to forecast the group type with high accuracy. In addition, we present a comparison between declared and detected groups along topicality and sociality dimensions.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services; H.1.2 [**User/Machine Systems**]: Human information processing;

---

[*]P.A.G. has performed the research while visiting Yahoo! Research Barcelona.

General Terms

Algorithms, Experimentation, Measurement

Keywords

Social Media, Groups, Bond theory, Identity theory, Flickr

1. INTRODUCTION

A longstanding theory about the creation of social communities affirms that people join groups driven by either strong personal ties with other members or by the interest in the group as a whole. As a result, depending on the prevalent motivation of members, spontaneously generated groups can be categorized as either *social* or *topical*. This theory is known as "common identity and common bond" [22] and assumes that the two types of groups have distinct and well recognizable traits that characterize them.

In recent years, the theory has been widely commented and elaborated by social scientists from a theoretical perspective and through small-scale experiments [26, 29, 24], but a validation over large-scale datasets together with the development of rigorous, automated methodologies to distinguish the group types is missing. Indeed, the availability of big data from social media platforms provides the opportunity to study the dynamics of the online communities from a data-mining perspective [19, 21, 14]. None of those experiments, however, have been directly aimed at verifying the common identity and common bond theory.

The design of a technique to detect the group type based on the common identity and common bond principles would first contribute to a strong validation of the theory itself and, on the other hand, would provide a general framework for automatic classification of user groups in online social media. Such a classification would have direct impact on the ability of describing the structure of online social networks along the axes of sociality and topicality and, consequently, on the possibility of better user characterization.

We contribute to fill this gap by proposing a set of general metrics based on the theory. We show that the metrics' values computed on a large corpus of groups extracted from Flickr confirm the cardinal points of the theory and are indeed good predictors of the group type. In addition, we repeat the same analysis on groups identified by a graph-based community detection algorithm. This allows us to compare the user-generated communities to the automatically detected ones not only from a structural perspective

but also along the dimensions of sociality and topicality. Since community detection techniques have been largely employed in recent years to describe the structure of complex social systems [8], the need for a clearer assessment of the meaning of the detected clusters has been often expressed from different angles [15, 31], but never completely satisfied. With our study we also contribute to shed light on this matter.

To the best of our knowledge, this is the first attempt of formalization of the common identity and common bond theory, and of its validation over a large and diverse set of user communities. The obtained results can open a new perspective on the semantic interpretation of implicit and explicit user groups in social media.

Our main contributions can be summarized as follows:

- Translation of the common identity and common bond theory into general metrics applicable to social graphs. An insightful characterization of a large group dataset from Flickr is performed using the proposed metrics.

- Comparison between user-defined groups and groups discovered by a community detection algorithm, both in terms of their overlap and their properties of sociality and topicality.

- Design of a method to predict whether a group is social or topical, based on the defined metrics. Prediction on the user-generated groups from the Flickr data yields surprisingly good results.

The rest of the paper is structured as follows. In Section 2 we overview related work. In Section 3 we describe the main principles of the common identity and common bond theory and propose their practical translation into general metrics, and in Section 4 we introduce the Flickr dataset we use to test these metrics. In Section 5 we describe the process of gathering a ground truth about group topicality or sociality through an editorial process. In Section 6 we provide a comparison between user-created Flickr groups and groups identified by a community detection algorithm in terms of their overlap, then we study the properties of the defined metrics computed over declared and detected groups and in particular over the subsets of groups labeled as social or topical in the ground truth that we produced. Finally, in Section 7 we show that we can accurately predict the group type by combining the defined metrics.

2. RELATED WORK

The spontaneous aggregation of people in communities has been widely studied in social science as one of the fundamental processes driving the global dynamics of social ecosystems [18] and as a factor driving the formation of the social identity of individuals [27].

Since the emergence of online social media, the global structure, evolution and dynamics of social groups have been investigated over large-scale and heterogeneous datasets [12]. Evolution of groups has been characterized as a very broad phenomenon [19, 5] that is dependent on the nature of the group [6], its intrinsic fitness [11] and on the density of social ties connecting its members [2]. Dependency of activity and connectivity on group size has been studied in several platforms [13, 14, 10], showing relations to Dunbar's theory on the upper bound of around 150 stable social relationships for an average human [7]. Besides activity, similarity between users is an important dimension in modeling individual users in groups [28], particularly given that, to a large extent, social media users tend to aggregate following the homophily principle [1]. Nevertheless, similarity is not necessarily the best indicator for group activity and longevity, as diversity of content shared between group members is a relevant factor to keep alive the interest of members [17].

At a finer scale, social communities can be described in terms of user engagement. From a quantitative perspective, the amount of participation of members in activities related to the group is varied and dependent on group size [3]. Intra-group activity has been characterized in terms of propensity of people to reply to questions of other members [30], coherence of discussion topics [9], or item sharing practices [21]. Modeling inner activity of groups has helped in finding effective strategies to predict future group growth or activity [14], recommend group affiliation, or enhance the search experience on social platforms [20].

Social and thematic components of communities have been widely studied in social science, most of all within the common identity and common bond theory on which the present work is based [22, 26, 24]. Nevertheless, the principles behind the theory have never been translated into practical methods to categorize groups, nor tested on large datasets. On the other hand, data-driven studies have investigated social and thematic components separately when characterizing groups [5]. Preliminary insights on the interweavement between such dimensions have been given in exploratory work on Flickr, where signals of correlation between social density and tag dispersion in groups is shown [23]. In this paper, we go far beyond that point, defining metrics that can be used to predict if a group is social or topical and testing their effectiveness against a reliable ground truth.

Besides the analysis of user-created groups, the study of automatically detected groups through community detection algorithms has attracted much interest lately [8]. Detected communities are supposed to represent meaningful aggregations of people where dense or intense social exchanges take place among members [13]. Nevertheless, even if synthetic methods to verify the quality of clusters have been proposed [15], the question of whether such artificial groups capture some notion of community perceived by the users remains open. If on the one hand the computation of cluster-goodness metrics over user-created groups can give useful hints about their structural cohesion [31], 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.

3. TOPICAL AND SOCIAL GROUPS

Notions of *community* and social *group* have been widely studied in the behavioral sciences [25, 27]. It has been shown that the internal dynamics of social groups emerge from the combination of complex cognitive processes such as sense of membership, influence between people, fulfillment of individual and collective needs inside the group, and shared emotional connections [18]. Based on such widely accepted theoretical foundations, sociological theories have been formulated to disentangle all of these complex aspects.

In the following subsections, we provide a high-level description of one of the most known theories on group forma-

tion and propose a translation of its main cornerstones into general metrics that can be applied on social graphs.

3.1 Identity and Bond Theory

The *common identity and common bond theory* describes social groups along the dimensions of topicality and sociality [22, 24]. According to the theory, the attachment to a group, as well as the permanence and involvement in it, can be explained in terms of common *identity* or common *bond*. Identity-based attachment holds when people join a group based on their interest in the community as a whole or in a well-defined common theme shared by all of the members. People whose participation is due to identity-based attachment may not directly engage with anyone and might even participate anonymously. Conversely, bond-based attachment is driven by personal social relations with other specific members, and thus the main theme of the group may be disregarded. The two processes result in two different group types, that for simplicity we name "*topical*" for identity-based attachment and "*social*" for bond-based attachment.

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

3.2 From theory to metrics

Based on the theoretical principles described above, it is possible to construct metrics to differentiate between the two types of groups. In particular, it is possible to quantify the reciprocity of interactions, and the topicality of the information exchanged between group members. We adopt a very general *multidigraph* model that fits most of the current social media platforms. Members are represented as nodes, and each distinct *interaction* between any two members is represented by a directed arc. Nodes can belong to multiple *groups* and we associate, with each group, a bag of user-generated *terms* (e.g., tags, group posts).

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.

3.2.1 Reciprocity

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

$$r_g^{\text{int}} = \frac{E_g^{\text{int,rec}}/2}{E_g^{\text{int,rec}}/2 + E_g^{\text{int,nrec}}}, \tag{1}$$

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

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

$$t_g = \frac{r_g^{\text{int}}}{\langle r_g^{\text{int}} \rangle}. \tag{2}$$

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

$$u_g = \frac{r_g^{\text{int}} + 1}{r_g^{\text{ext}} + 1}. \tag{3}$$

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

3.2.2 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

$$H(g) = - \sum_{t \in T(g)} p(t) \cdot \log_2 p(t), \tag{4}$$

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:

$$h_g = \frac{H(g)}{\langle H(f) \rangle_{|T(g)|=|T(f)|}}. \tag{5}$$

3.2.3 Activity

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

$$a_g = \frac{E_g^{\text{int}}}{(D_g^{\text{in}} D_g^{\text{out}})/E}. \tag{6}$$

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

comments	favorites	contacts	decl. g.	det. g.
238M	112M	71M	504K	646K

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

this group, where E is the total number of interactions in the network. If this property has a value higher than 1 then the number of interactions internal to the group is higher than the number of interactions expected in a random scenario with the same group activity volume.

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

$$b_g = \frac{E_g^{\text{int}}/(s_g(s_g-1))}{E_g^{\text{ext}}/(2(N-s_g)s_g)}, \qquad (7)$$

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

4. DATASET AND PREPROCESSING

The wide variety of user groups and the richness of interaction types make Flickr an ideal platform for our study. We use only public, anonymous data retrievable via the Flickr public API, until the end of 2008. Table 1 summarizes the data described below.

4.1 User interactions

We collected three types of pairwise, directed 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.

4.2 Groups

Users of Flickr can create, moderate and administer their own groups. Most groups are open, so users can join without an invitation. Others are only by invitation and joining requires the administrator's permission. There are over 500K groups in our Flickr dataset.

In addition to user-created groups (we refer to them as *declared*), we analyze the sociality and topicality properties of groups that are not defined by users but are instead found by community detection algorithms (we name these *detected* groups). We applied the OSLOM community detection algorithm [16] 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 [15] and it outperformed other algorithms we tested. OSLOM detected 646K groups.

4.3 Tags

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

5. GROUP LABELING

To determine whether the defined metrics correctly capture the sociality and topicality of groups, we compare them against a reliable ground truth. We asked human editors to label groups based on well-defined guidelines extracted directly from the common identity and common bond theory [24]. For the labeling we randomly selected groups meeting the following requirements: i) more than 5 members, ii) more than 100 internal comments, iii) relative activities a_g^{com} and b_g^{com} higher than 10^2. The third requirement ensured us that the selected groups were active well above the expected values in a random case. After this selection we obtained over 34K declared groups and over 33K detected groups. We describe the labeling process of such groups in detail in the following subsections.

5.1 Information provided to editors

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

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

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

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

5.2 Labeling guidelines

Human labelers were shown the information described above and asked to categorize groups as either *social, topical* or *unknown.* The last case is reserved for groups for which text is written in a language unknown to the labeler, making the task impossible to accomplish. Intentionally, no *unsure* category was allowed to keep the categorization strictly binary, as the theory does. Some groups can be both topical and social, and therefore difficult to categorize, but for the

sake of clarity and conformity with the theory we kept the categorization as a binary task. Editors were provided with specific instructions on how to recognize social and topical groups, and on how to perform the categorization. The guidelines are summarized as follows:

I. Comments and photos. By examining comments and photos, find traces of people who know each other or who have a personal relationship. Knowing each other's real names, spending time together, co-appearing in photos, sharing common past experiences, referencing mutually known places, and disclosing personal information are all signals of the presence of a social relationship [4]. The predominance of friendly and colloquial comments (e.g., jokes, laughter) is another element distinguishing social groups from topical groups. In topical groups, the atmosphere is more formal and comments tend to be more impersonal [26]. Examples of impersonal comments include expressing appreciation for photos, praising the photographers, thanking them for their work, or commenting on any particular topic in a neutral way. As a rule of thumb, if many personal comments are detected, then the sociality of the group should be considered high. If such comments are not many (e.g., just between small subsets of members), but the overall atmosphere of the interaction is rather personal and friendly, then we consider the sociality of this group as fairly present. If, on the other hand, comments are mainly impersonal and neutral, sociality has to be considered low, in favor of higher topicality.

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

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

5.3 Group examples

To provide a sense of how the defined guidelines were applied in practice, we describe two examples. The first one is a group titled "Airlines Austrian", tagged with labels "aircraft", "airport" and "spotting." Photos are from different countries in Europe and the vast majority of them depict airplanes. Members are very active in commenting and writing comments related on the aircraft theme (e.g., "I just love this airplane, the TU-154M is just a plane Boeing or Airbus could never design"). In this case, all of the features are aligned with the concept of topical group defined in the guidelines. The second group is named "Camp Baby 2008" and it is described in the main page as a collection of photos of a two-day event for young mothers taking place at a spe-
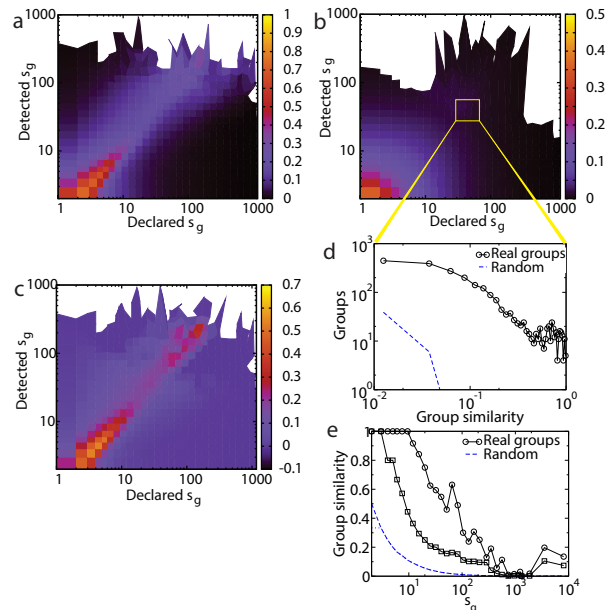

Figure 1: Jaccard similarity between declared and detected groups as a function of their sizes. Diagonal shows an interesting pattern (a) which is not reproduced by randomized groups (b). We subtract (b) from (a) and plot the result in (c), and a histogram of similarity values for a sample of groups lying at the diagonal in (d). For groups of various sizes, we plot 91th and 99th percentiles of similarity between declared and detected groups (e).

cific location. Photos depict people attending the event and interacting with each other with a friendly attitude. Tags and comments often contain names of individuals and references to past common experiences (e.g., "I love Mindy and can not wait to see her again!!"). Although the group has a specific topic, its social component is very strong.

In practice, more ambiguous cases can occur and, ultimately, the decision of the labeler has an arbitrary component, as in every complex annotation process. Nevertheless, the defined guidelines gave the labelers precise instructions and, as described next, we recurred to multiple independent editors to assess the quality of the extracted ground truth.

5.4 Labeling outcome

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

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

The Kappa value for detected groups is around 0.44, revealing lower agreement. A factor that partially determined
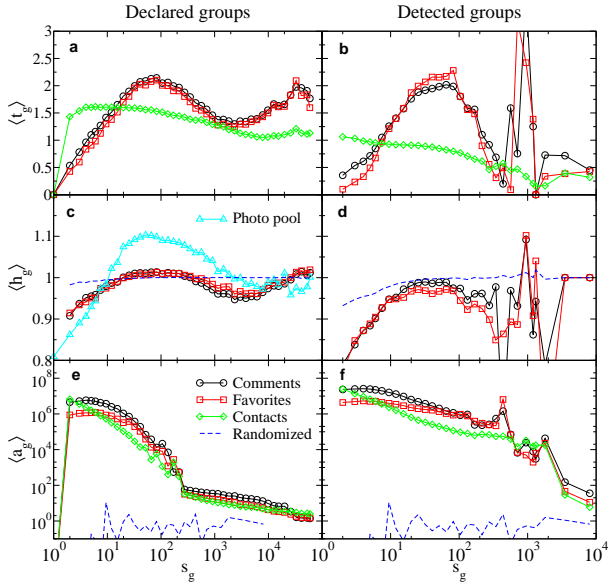
Figure 2: Dependency of normalized reciprocity (a-b), normalized entropy (c-d) and relative activity (e-f) on size of groups for comments, favorites, contacts and photo pools, for declared and detected groups. Blue dashed line is for randomized photos (c-d) and groups (e-f).

such result is the lack of information about the group's profile, since it is not available for detected groups. Another cause of the disagreement is a higher variability in the comments. This may be because we use a network of contacts for the purpose of finding clusters and defining detected groups, which may not be the best proxy of personal relations.

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

6. CHARACTERIZATION OF GROUPS

We begin the analysis with a direct comparison of the overlap between the declared and detected groups. Then we characterize the two sets of groups in terms of the metrics we introduced in Section 3.2. Finally, we study the relation between the labels of the declared groups annotated by the editors and the values of the metrics. Additionally, we report ratios of groups labeled as social and topical among both declared, and detected groups.

6.1 Membership overlap of declared and detected groups

The groups from the two sets share typical properties of groups found in on-line social networks. The distribution of sizes of groups in both cases is heavy-tailed and close to power-laws (not shown due to space limits). 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
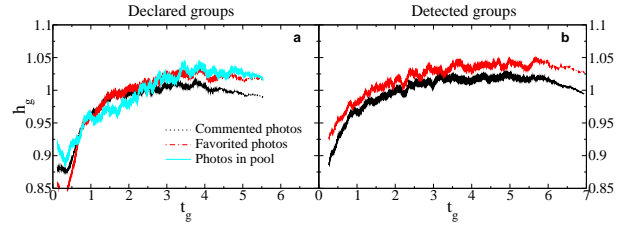

Figure 3: Correlation between reciprocity of comments inside a group and entropy of photos commented or favorited between its members, or belonging to the photo pool of this group, for declared and detected groups.

as a function of the size of groups in Fig. 1a. Zero values of similarity are not taken into account for these averages. For the purpose of comparison with a null model, in Fig. 1b 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. 1c, meaning that indeed detected groups are, to some extent, similar to the declared ones. Further insights are shown in Fig. 1d, where we depict the distribution of similarities of pairs of groups extracted from a small sector of the diagonal, having between 32 and 64 members. The figure shows that there exist multiple detected groups which overlap significantly with declared groups, and that randomized groups do not show this pattern. This holds for groups of all sizes, as shown in Fig. 1e, in which we plot the 91th and 99th percentiles of the best match similarity for detected groups of various sizes (e.g., 1% of detected groups of size 20 have similarity with declared groups higher than 0.75, while for the randomized case 1% of the groups have similarity higher than just 0.05). Therefore, in some cases the community detection algorithm finds groups which are also defined by users (i.e., declared groups). We present evidences that this does not occur by chance through the comparison with the randomized case. Nevertheless, a substantial overlap is found for just a small percentage of groups. Most of the group pairs have similarity close to 0. Consequently, the similarity of detected groups to the best-matching declared groups is 0.082, while for the randomized detected groups it is not much lower, yielding 0.058.

6.2 Statistical properties of metrics

Besides directly comparing membership overlap, we study the variation of the metrics defined in Section 3.2 with the group size. Reciprocity and normalized entropy have a wide local maximum for groups of sizes between 50 and 100 members, both for declared and detected groups, as shown in Figs. 2a-d. This holds for all interactions and all sets of tags, with the exception of contacts, for which the curves are relatively flat. A similar local maximum has been found in a recent study [13] for pairwise interactions in Twitter by various community detection algorithms. We perform a randomization of photos between groups, keeping the number of photos per group fixed. The normalized entropy calculated for the shuffled photos stays close to 1, as expected, and the maximum disappears. A possible interpretation of the existence of the maximum is that these sizes tend to correspond to social groups, while bigger groups are more frequently
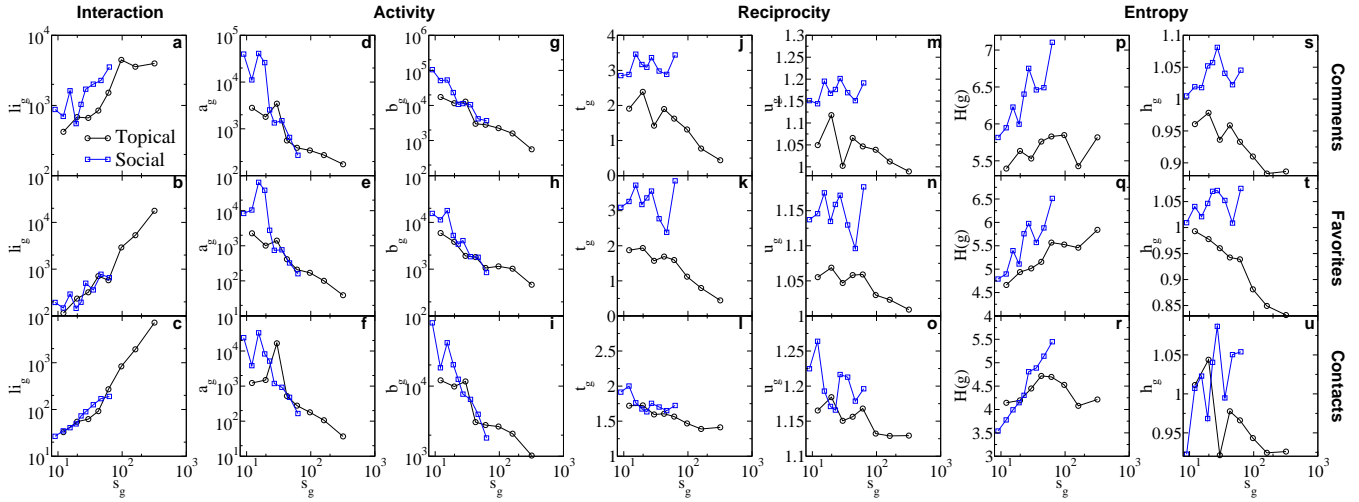
Figure 4: Averages of various properties of topical (black circles) and social (blue squares) groups as a function of their size. Each point corresponds to 30 groups.

topical. Further findings to support this interpretation are presented in the next subsection.

Strong correspondence of the maxima for normalized entropy and reciprocity suggests that these properties are correlated, as shown in Fig. 3. While it may be natural to explain the correlation between reciprocity of comments and normalized entropy based on commented photos, it is not clear why we also find a positive correlation with normalized entropy based on other sets of photos. A possible interpretation is that high intra-reciprocity leads to wider variety of topics covered inside of that group, and vice versa.

The values of relative activity both in declared and detected groups are very high, as presented in Figs. 2e,f. As expected, activity of randomized groups exhibits values around 1 for all group sizes. For real groups instead, the value of relative activity decreases with the size of groups and gets close to 1 for very large ones. This may be caused by the fact that larger groups cannot be as integrated as smaller groups and the social commitment of their members towards other members of the group drops due to limited human capabilities. Additionally, we observe that the activity decay for declared groups occurs sharply between groups of size 100 and 200, in agreement with Dunbar's theory on the upper bound of the number of stable relationships manageable by a human. The activity drop for detected groups is continuous and much more moderate (Fig. 2f), since community detection algorithms tend by design to output node clusters with high numbers of connections between them.

6.3   Relation between metrics and group label

Here we analyze properties and values of the metrics for groups labeled through the editorial process. First, the ratio of groups labeled as social differs between declared and detected groups. In declared groups we find around 48% social groups, whereas among detected groups almost 69% are labeled as social. Additionally, we picked 50 detected groups among the ones that are the most similar to declared groups. Specifically, we selected them randomly from the 99th percentile shown in Fig. 1. These groups have significant overlap with declared groups and should share similar properties. Indeed, the ratio of groups labeled as social

among them is closer to that of declared groups and equal to 53%. We conclude that detected groups are more likely to be social than declared ones. It is a somewhat expected result, since clustering algorithms detect dense parts of a network, and so they are inclined to detect areas with more reciprocal connections. Note that the theory envisions more reciprocal relations in social groups. Thus, community detection algorithms are more likely to find social groups, however, determining to what extent it happens is not trivial.

One of the expectations is that bond-based groups should not be very large, as the human capacity for stable relationships is limited. As pointed in Section 6.2, the Dunbar number can be considered as a possible cap for the size of such groups, while topical groups do no yield such a restriction. In line with this expectation, we find that declared groups labeled as social have on average 35 members, whereas groups labeled as topical have on average around 172 members.

We find insightful differences and similarities in various properties, which we explore in detail in Fig. 4. We plot them as a function of the size of groups as they vary drastically with it, and one needs to compare groups of similar sizes in order to draw unbiased conclusions.

First, there are almost no differences in the number of photos (not shown), favorites, and contacts (as in Figs. 4b,c) inside social and topical groups. The number of comments is, however, around 2 times higher in social groups than in topical groups of similar size (Fig. 4a). More differences can be found when looking at relative activity metrics (Figs. 4d-i), which compare number of internal interactions with an expectation. In all three types of interaction the relative activity metrics for social groups yield values from 2 to over 10 times higher than for topical groups. The activity metric b_g compares density of interactions internal to the group with the density external to it. Therefore this result reflects a stronger focus or even an isolation of members belonging to social groups from the rest of people they interact with.

More importantly, we observe large differences in values of reciprocity and relative reciprocity of comments and favorites. Social groups exhibit significantly higher reciprocity than topical groups (Figs. 4j-o), in line with common identity and common bond theory. There is no difference in
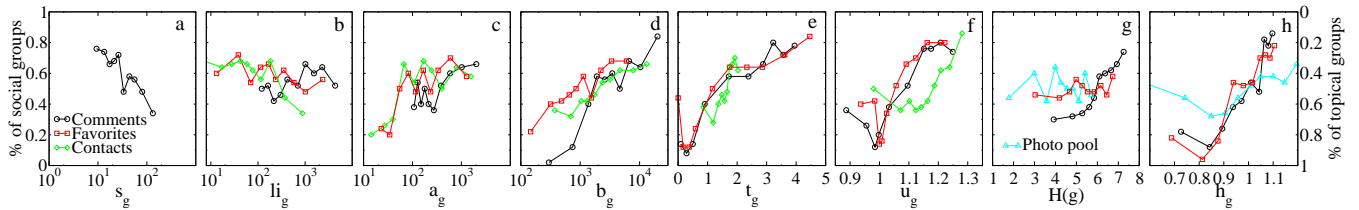
Figure 5: Dependence of fraction f of groups labeled as social on various metrics: based on comments, favorites, contacts and photo pools. The remaining $(1-f)$ groups are topical. Each point corresponds to 50 groups.

reciprocity of contacts. A plausible interpretation is that contacts do not reflect personal relations between connected users, as users often add people they do not know and do not interact with as contacts in order to follow their content. Finally, we observe much higher values of entropy and normalized entropy in social groups than in topical ones (Figs. 4p,q,s,t). This holds for the tags extracted from photos commented, and favorited between members. Assuming that tags of photos represent topics of interaction, the result is consistent with bond attachment. It is expected for members of bond-based groups to cover many different topics and areas in their interactions, whereas members of identity-based groups focus their interactions on specific topics. However, this effect is weaker for the tags extracted from photo pool of the group (Figs. 4r,u). Apparently, the content of the photo pool does not always reflect well the interactions and relations between members of the group.

Additionally, we plot the fraction of groups labeled as social with respect to group size, activity, reciprocity, and entropy (Fig. 5). The fraction correlates negatively with group size, as expected (Fig. 5a). The correlations with the number of interactions and relative activity a_g are rather weak (Figs. 5b,c), whereas, surprisingly, there is a strong dependency on relative activity b_g (Fig. 5d). For the lowest values of b_g^{com}, 95% of the groups are topical, while for the highest, 80% of the groups are social. High values of b_g can mean stronger group-focus, or even an isolation of the group members from the rest of people they interact with. It may relate to the fact that it is hard to enter bond-based groups due to strong relations existing between their members and because high investment is required to create such relations with them [24]. Direct reciprocity of interactions, with the exception of contacts, correlates strongly with social groups (Figs. 5e,f). We strongly expected this result based on bond attachment. Furthermore, we found that the entropy of tags correlates with social groups, but entropy based on other sources does not (Fig. 5g). However, we find that our normalized entropy performs much better than this, and a strong correlation is found both for tags extracted from comments and from favorites (Fig. 5h). This shows that the normalized entropy of tags may be the most proper way of measuring topical diversity of communications of a set of people.

7. GROUP TYPE DETECTION

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

7.1 Prediction methodology

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

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

7.2 Prediction results

The ratio of groups labeled as social increases quickly with the score S_g, as shown in Fig. 6a. This summarizes the findings of previous sections, suggesting that the features embedded in the score are able to capture well the nature of the groups. The higher the score, the higher the probability that the group is social; the lower, the more topical. When the score is around zero, groups can be either social or top-
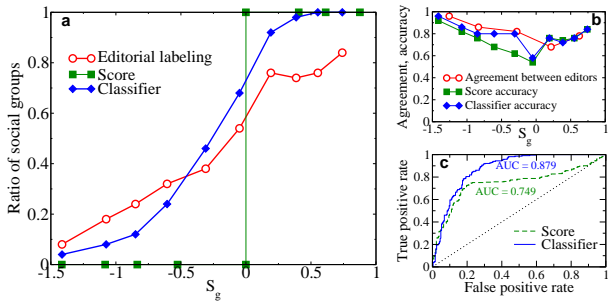
Figure 6: Comparison between prediction methods. a) Ratio of groups labeled as social versus the score. The same ratio is plotted for the score-based (threshold at 0) and trained classifier predictions. b) The accuracy of prediction of the two techniques and agreement between two of the labelers against the S_g values. c) ROC curves for the prediction with the two different techniques.

ical, or both, meaning that a decision on the nature of the group may be more difficult. If we fix the threshold for the S_g value in order to perform a binary group classification, it is clear that several misclassifications will occur, especially around the threshold value. An example for threshold at 0 is shown in Fig. 6a. 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. 6b). The accuracy at the extreme values of the score is close to 0.95, while it falls below 0.6 for groups with a score close to 0. On the other hand, this drop appears also in the agreement between two of the human labelers, measured as a ratio of groups that have been given the same label. Apparently, this is a shortcoming of the binary classification itself, as opposed to multi-label classification.

The overall performance of the two approaches can be compared fairly through ROC curves (Fig. 6c), 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 Tab. 2.

In addition, to determine the most predictive features, we rank the features using Chi-square feature selection. The top 5 features are, in decreasing order of importance: h_g^{com}, t_g^{com}, u_g^{com}, h_g^{fav}, and b_g^{com}. The selected set is the optimal for the prediction performance: retraining the classifier on such restricted set of features results in stable performance, as shown in Tab. 2. The top 4 most predictive features correspond directly to the expectations of the theory and results of the analysis from Section 6. Reciprocity-based metrics and normalized entropy are significantly more predictive than other features. The high position of relative activity b_g^{com} is rather unexpected. However, we have already remarked on its importance and commented on its interpretation in Section 6.

8. CONCLUSIONS

Common identity and common bond theory indicates a

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

Table 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.

high-level characterization of topical and social groups. We propose metrics capturing reciprocity of interactions and entropy of user-generated terms, to realize the concepts discussed in the theory and to measure sociality and topicality of groups. We label a number of groups from Flickr as either topical or social and we leverage this ground truth to show that the metrics, combined with a machine-learning approach, predict the group type with high accuracy. Moreover, we note that the degree of isolation of the group activity from the rest of the social network, measured in terms of more personal interactions, is a good predictor of the group type, in addition to the elements identified in the theory. Besides the main prediction results, the supporting analysis of the group properties in terms of the identified dimensions confirms the theory from different angles and highlights other interesting findings. In particular, dependencies of the metrics with the group size confirm previous observations about the effective size of social communities, peaking around rather small sizes and being limited by a cap of 100-200 members.

The study is complemented with a comparison of the structure and sociality and topicality traits between declared groups and groups from community detection algorithms. Detected groups do not overlap much with declared groups on average, but they match sensibly more than the random case for groups of comparable sizes. Furthermore, detected groups are more often social than the declared ones.

Findings from the present work open new opportunities for characterization of social communities and of their members. Extensions to the study include a more exhaustive extraction of detected groups using a different network than the network of contacts e.g., we find mutual comments to carry more social traits than the contacts do. Another interesting extension could be multi-label classification of groups, in order to better categorize groups with mixed social and topical components.

Acknowledgments

9. REFERENCES

[1] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. Friendship prediction and homophily in social media. *ACM Trans. Web*, 6(2):9:1–9:33, 2012.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD'06*, page 44, New York, New York, USA, 2006.

[3] L. Backstrom, R. Kumar, C. Marlow, J. Novak, and A. Tomkins. Preferential behavior in online groups. In *International conference on Web search and web data mining - WSDM'08*, pages 117–128, New York, NY, USA, 2008.

[4] N. L. Collins and L. C. Miller. Self-disclosure and liking: A meta-analytic review. *Psychological Bulletin*, 166(3):457–475, 1994.

[5] A. Cox, P. Clough, and S. Siersdorfer. Developing metrics to characterize Flickr groups. *Journal of the American Society for Information Science and Technology*, 62(1):493–506, 2011.

[6] J. N. Cummings, B. Butler, and R. Kraut. The quality of online social relationships. *Comm. of the ACM*, 45(7):103–108, 2002.

[7] R. I. M. Dunbar. The social brain hypothesis. *Evolutionary Anthropology*, 6:178–190, 1998.

[8] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[9] P. A. Gloor and Y. Zhao. Analyzing Actors and Their Discussion Topics by Semantic Social Network Analysis. In *Conference on Information Visualization*, IV '06, pages 130–135, Washington, DC, USA, 2006.

[10] B. Goncalves, N. Perra, and A. Vespignani. Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number. *PLoS ONE*, 6(8):e22656, 2011.

[11] P. A. Grabowicz and V. M. Eguíluz. Heterogeneity shapes groups growth in social online communities. *Europhys. Lett.*, 97(2):28002, 2012.

[12] P. A. Grabowicz, J. J. Ramasco, and V. M. Eguiluz. Dynamics in online social networks. *arXiv*:1210.0808, 2012.

[13] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz. Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media. *PLoS ONE*, 7(1):e29358, 2012.

[14] S. R. Kairam, D. J. Wang, and J. Leskovec. The life and death of online groups: predicting group growth and longevity. In *5th ACM international conference on Web search and data mining*, WSDM'12, pages 673–682, New York, NY, USA, 2012.

[15] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78:046110, 2008.

[16] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato. Finding Statistically Significant Communities in Networks. *PLoS ONE*, 6(4):e18961, 2011.

[17] P. J. Ludford, D. Cosley, D. Frankowski, and L. Terveen. Think different: increasing online community participation using uniqueness and group dissimilarity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 631–638, New York, NY, USA, 2004.

[18] D. W. McMillan and D. M. Chavis. Sense of community: A definition and theory. *Journal of Community Psychology*, 14(1):6–23, 1986.

[19] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *7th ACM SIGCOMM conference on Internet measurement - IMC '07*, pages 29–42, San Diego, California, USA, 2007.

[20] R.-A. Negoescu, B. Adams, D. Phung, S. Venkatesh, and D. Gatica-Perez. Flickr hypergroups. In *17th ACM international conference on Multimedia*, MM'09, pages 813–816, New York, NY, USA, 2009.

[21] R. A. Negoescu and D. Gatica-Perez. Analyzing flickr groups. In *International conference on Content-based image and video retrieval*, CIVR'08, pages 417–426, New York, NY, USA, 2008.

[22] D. A. Prentice, D. T. Miller, and J. R. Lightdale. Asymmetries in attachments to groups and to their members: Distinguishing between common-identity and common-bond groups. *Personality and Social Psychology Bulletin*, 20(5):484–493, 1994.

[23] C. Prieur, N. Pissard, J. Beuscart, and D. Cardon. Thematic and social indicators for Flickr groups. In *Proceedings of International Conference on Weblogs and Social Media - ICWSM'08*, 2008.

[24] Y. Ren, R. Kraut, and S. Kiesler. Applying Common Identity and Bond Theory to Design of Online Communities. *Organization Studies*, 28(3):377–408, 2007.

[25] S. Riger and P. J. Lavrakas. Community ties: Patterns of attachment and social interaction in urban neighborhoods. *American Journal of Community Psychology*, 9:55–66, 1981.

[26] K. Sassenberg. Common bond and common identity groups on the Internet: Attachment and normative behavior in on-topic and off-topic chats. *Group Dynamics Theory Research And Practice*, 6(1):27–37, 2002.

[27] H. Tajfel. *Social identity and intergroup relations*. Cambridge University Press, 1982.

[28] L. Tang, X. Wang, and H. Liu. Group profiling for understanding social structures. *ACM Trans. Intell. Syst. Technol.*, 3(1):15:1–15:25, 2011.

[29] S. Utz and K. Sassenberg. Distributive justice in common-bond and common-identity groups. *Group Processes and Intergroup Relations*, 5(2):151–162, 2002.

[30] H. T. Welser, E. Gleave, D. Fisher, and M. Smith. Visualizing the Signatures of Social Roles in Online Discussion Groups. *The Journal of Social Structure*, 8(2), 2007.

[31] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS'12, pages 3:1–3:8, New York, NY, USA, 2012.