

Developing Metrics to Characterize Flickr Groups

Andrew Cox and Paul Clough

The Information School, University of Sheffield, Room 226, Regent Court, 211 Portobello Street, Sheffield, S1 4DP UK. E-mail: {a.m.cox, p.d.clough}@sheffield.ac.uk

Stefan Siersdorfer

L3S Research Center, Appelstr. 9a 30167, Hannover, Germany. E-mail: siersdorfer@l3s.de

Flickr, the large-scale online photo sharing website, is often viewed as one of the ‘classic’ examples of Web2.0 applications through which researchers are able to observe the social behavior of online communities. One of the main features of Flickr is groups. These provide a means to organize, share and discuss photos of potential interest to group members. This paper explores the scale of group creation on Flickr and proposes a new set of metrics for characterizing groups on Flickr looking at aspects of membership, communication activity, and communication structure. Data collected from a sample of 1,000 groups was used to confirm the metrics and provide new insights into group formation in Flickr, such as the nature of larger and smaller groups. The contributions of the article are as follows: a set of metrics for characterizing online groups that extend existing schemes; an approach for sampling Flickr to estimate the number of groups; new insights into Flickr groups based on results from analyzing 1,000 randomly selected groups; and reflections on our experiences with using publicly accessible, automatically collected data to characterize the types of groups on Flickr.

Introduction

Large Web2.0 sites and other “mega Web sites” like Flickr are difficult to grasp because of their sheer scale. In October 2009 Flickr claimed to have 4 billion photos,¹ having passed the 3 billion mark just 6 months earlier. The site may produce many rich personal experiences, but it is hard to represent objectively the features of the whole site. Individual impressions can be quite different because it is more like a place where many communities are supported than itself a community (Perez, 2007).

Received July 29, 2010; revised November 19, 2010; accepted December 11, 2010

¹<http://blog.flickr.net/2009/10/12/4000000000/> (site accessed: 26 July 2010)

© 2011 ASIS&T • Published online 25 January 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.21483

One of the most studied features of the Internet as a phenomenon is the active involvement of users in creating online or virtual communities. Usenet, listservs, and Web forums, such as Yahoo! Groups have supported the emergence of vibrant collectivities. Web2.0 sites, such as Flickr and Facebook, however, seem to be based primarily on a more social network model. Thus, Flickr would be understood via individual users uploading and organizing personal collections of photos, making social contacts through browsing profiles, and responding to comments. “People define their community egocentrically” (boyd 2006). Yet Flickr has a group function that is heavily used. To what extent should Flickr groups be understood through the online community model? Is part of the success of Flickr its combination of ego-centred activity with the online community concept?

Flickr is a large site based on user generated photos and often cited as a “classic” example of Web2.0 (Cox, 2008; van House et al., 2005). Flickr has been the object of many studies, including attempts to characterize users (Miller & Edwards 2007; Cox et al., 2008), the tags users assign to photos (Rorissa, 2010), and investigating user’s motivations for publishing and tagging (Nov et al., 2008; Angus & Thelwall, 2010), for example. Other studies have looked at how Flickr data can be used for a variety of purposes, such as providing recommendations for tagging photos (Sigurbjörnsson & van Zwol, 2008), and automatically assigning geographic coordinates to Flickr photos (Van Laere et al., 2010). Some fascinating Flickr group related tools have been designed, such as Group trackr.²

Flickr groups³ have a pool of photos, discussion space, and a list of members. Anyone can set up a group, choose a name, define some simple functions, and become an “Admin.” There are private and public groups. Generally, users can join a group just by clicking on a button. They can then choose to

²http://dev.nitens.org/flickr/group_trackr.php (site accessed: 26 July 2010)

³FAQ for Flickr groups: <http://www.flickr.com/help/groups/> (site accessed: 26 July 2010)

surface one of their photos in the group pool. Several different types of groups seem to have emerged on Flickr (Davies, 2006; Stvilia & Jørgensen, 2009; Malinen, 2010). Some are places to collect photos together around a particular topic, theme, or photographic technique, be this very specific, such as pictures of a type of tree, or something very broad, such as black and white photos. Given that tagging solves only partly the problem of organizing the vast collection of photos on Flickr, adding photos to groups is a form of emergent topical classification. Other groups are organized to hold competitions or make awards, effectively highlighting good photos. These groups can be themselves thematic or very broad in scope. How the awards are made varies, but often the group Admin and helpers choose whom to reward. The existence of award groups is linked to the desire to identify good photos, again, in an emergent way. An interestingness algorithm is Flickr's own way of calculating how good a photo is, which uses evidence such as the number of times a photo has been viewed, commented, or favorited to priorities photos in searches. The award groups are another bottom-up approach to organizing photos by quality, which allows many different criteria of goodness to be coexisting on Flickr. A third type of Flickr group has quite a strong geographical basis, collecting photos of a particular town or village, and these seem to be the most likely to organize meet-ups and develop into social groups.

On a superficial level, Flickr groups resemble the organization of Yahoo! groups. In fact, a comparison of the organization of Flickr and Yahoo! groups is instructive. Whereas Yahoo! groups are relatively clearly bounded and can be browsed as separate entities, because of the rich navigation paths, it is very easy within Flickr to navigate out of any Flickr group, e.g., from a photo in the pool to the "photostream" of its author or from the membership list to the profile of individual members, their photos, favorites, or contacts. There is no function to browse Flickr groups. Actually, Flickr is far less organized around groups; rather, it is focussed on individuals, their uploading and organizing of their own photos, monitoring traffic on these photos, and managing their own navigation of the site by recording photos, people, and groups they like (via favorites, contacts, and group memberships). Data based on reciprocal commenting and friending would be the most direct way to explore the social network (Recuero, 2007; Prieur et al., 2008), perhaps using social network analysis combined with a sense of community questionnaire as proposed for blogging by Chin and Chigwell (2006). However, the groups exist and clearly play some function within Flickr as a very successful and engaging site. Given the scale of Flickr, it is useful to try and use some quantitative approaches to understanding the role of groups. This article provides the foundations for such an approach by the following research objectives: first, to establish the scale of group creation on the site; second, to develop a set of metrics that use publicly accessible, automatically collected data to characterize groups in Flickr; and third, to use these metrics to characterize the scale and activity Flickr groups and the types of community it supports. The Related Work section

describes past literature related to studying Flickr groups and online communities. The Proposed Metrics section presents our proposed metrics for characterizing online groups, which we use to analyse Flickr groups. The Methodology section describes our methodology used to gather data for analyzing Flickr groups. The results are presented in the Results and Analysis section, and a discussion follows in the Discussion section. Finally, in the Conclusions and Future Work section, we conclude the article and discuss avenues for further research.

Related Work

There has been relatively little academic research on Flickr groups as such. A notable exception is the work of Negoescu and Gatica-Perez (2008). They looked at group behavior from a collection of the 500 most recent photos from 22,414 recently active users. They found that only around half of the sample had ever surfaced a photo in a group pool. That group activity is not universal is confirmed by Stvilia and Jørgensen (2009), who found that 38% of users sampled did not join groups and that the median number of members was only two. In Negoescu and Gatica-Perez's (2008) study, paying members of Flickr ("pros") put more photos in groups than nonpaying members, perhaps partly as a function of their having more photos. About a quarter had "shared" photos in over 50 groups; 10% in over 200 groups. Fifteen percent of the people who had put a photo in a group had shared all the photos they had shared with just one group; 45% had shared at least one photo with more than 20 groups. Thus, on the whole where people do surface photos in groups, they tend to use multiple groups. But the authors found that users tended not to share lots of photos with a particular group; "loyalty" was low. Eighty-five percent of users had an average of less than 15 photos in any particular group. On the other hand, the same photo is not put in lots of groups—there is not much "photo recycling"—; the average is about 3.1 groups per photo, implying that most users do not try and promote a photo on lots of groups, even though, as Sigurbjörnsson and van Zwol (2008, p. 5) found, there is a correlation between the number of times a photo is viewed and the number of groups it is in.

The measures that Negoescu and Gatica-Perez (2008) used in their investigation tell us something broadly about the average sort of behavior in groups, but they do not tell us very much about the character of groups themselves. To develop a fuller range of descriptive metrics for Flickr groups, a useful starting point is Butler's (1999) working paper on listservs (see also Cummings et al., 2002). Here, he is considering whether listservs should be best understood as analogous to small groups or voluntary associations, and he proposes sets of measures under the headings of membership, communication activity, and participation structure to explore this. His metrics are summarized in Table 1.

These metrics are well chosen to capture the basic scale and turnover of membership, the raw level of communication, and the character of participation, whether it seems to reflect reciprocation and what proportion of the membership

TABLE 1. Descriptive metrics for online collectives (Butler, 1999).

A. Membership
Size
Growth / loss / change
B. Communication activity
Average number of messages per day
Percentage of all groups with zero activity
C. Participation structure
Average thread length
Proportion of messages receiving no reply
Interactive or episodic pattern of messaging
Participation ratio—proportion of members who contribute at least one message
Gini coefficient for the distribution of participation among active participants
Proportion of all messages sent by two top participants

is involved in discussion. With some modification, the metrics also seem relevant to identifying the different types of Flickr groups. The membership measures are relevant, although it might be hypothesized that because the effort involved in being a member of a listserv is greater than joining a Flickr group—one has to do something with incoming e-mails—, Flickr groups might be expected to be larger and more stable because there is little incentive to leave a group on Flickr. Listservs function through the circulation of text messages; on Flickr, normally the key group activity is surfacing photos in the pool. Thus, the equivalent to the average number of messages would be the rate of posting of photos. In addition to photo-related activity, discussions on the Flickr group forum could also be considered in a way more obviously analogous to Butler’s metric. Participation structure is slightly more difficult to translate for Flickr. Butler’s first three structural measures are about interaction, the real equivalent of which would be responses from surfacing a photo in terms of viewing, favoriting, and commenting. Such activity is much more difficult to estimate in Flickr, partly because a photo may be found through a number of routes and so the level of viewing may not necessarily be linked to it being placed in a group pool. Thread length, etc., can be found for the group forum. The second set of three measures is about levels of participation and the distribution of activity across the whole group or a small core of users. Core activity is possible to calculate for Flickr, e.g., by estimating what proportion of all pool photos were posted by the most frequently posting members.

Butler’s (1999) work provides an excellent starting point for understanding appropriate group metrics, but, surprisingly, we can find few subsequent studies that extend these metrics for discussion forums or other types of site. Similar types of metric have been proposed for communities of practice (Castro, 2006) and for marketing communities (Cothrel, 2000). Schoberth et al. (2003) look at some of the same measures in relation to longitudinal group activity and Smith and colleagues have produced many metrics and visualizations for usenet in the Netscan project (e.g., Smith, 2002).

TABLE 2. Yahoo! groups metrics (Backstrom et al., 2008).

Base line traffic—2 messages in every monthly interval
Base line users—10 distinct users post in a year
Dense period —2-month-period in which every 7-day period there are 10 + posts
A membership core, based on replies to and replies from numbers of distinct other users

Another interesting approach is offered by Backstrom et al.’s (2008) study of Yahoo! groups. Some of their metrics are summarized in Table 2.

Whereas Butler is trying to characterize the communicative character of online collectives, Backstrom et al. (2008) are seeking to identify segments of the groups that are active, and so they propose critical levels of membership or activity that indicate the strength of the group. They segment Yahoo! groups with active small/private groups and less active, very large public groups. Although this does not add to Butler’s measures, it does suggest an approach to how to use the figures, i.e., to segment the whole population.

Whereas the quantitative study of online community has not been very fully developed, there have been many qualitative studies of online communities. Many of these have been motivated by a desire to explain what features of these groups make users feel them to be communities. As part of an attempt to operationalize these parameters for content or linguistic analysis, Herring (2004) usefully summarizes some of the key features of online groups that have been identified as indicating community. These are summarized by her in the quotation below:

- 1) Active, self-sustaining participation: a core of regular participants
- 2) Shared history, purpose, culture, norms and values
- 3) Solidarity, support, reciprocity
- 4) Criticism, conflict, means of conflict resolution
- 5) Self-awareness of group as an entity distinct from other groups
- 6) Emergence of rules, hierarchy, governance, rituals (Herring 2004)

Not surprisingly, these qualitative aspects of a group are hard to operationalize in a form whose collection can be automated. The approach, such as developed by Blanchard and Markus (2004; Blanchard 2007) based on the McMillan and Chavis (1986) definition of “sense of community,” is convincing but has to be questionnaire-based, and so it is dependent on response rates and difficult to scale to sites such as Flickr. However, some of the qualitative aspects can be operationalized in quantitative terms. Thus, Herring’s concept of self-sustaining participation does seem to relate closely to aspects of Butler’s (1999) participation structure, such as the participation ratio. Herring’s third point about solidarity relate to Butler’s metric about interactivity in messaging. We also suggest that for Flickr, some limited indication of the existence of the emergence of rules and hierarchy could be taken from the length of the group description text and

the number of administrators and moderators that have been defined for the group.

Proposed Metrics

From previous literature, we produced a preliminary set of metrics for Flickr groups before considering what was technically feasible. The purpose of such quantitative measures is to try and characterize the types of Flickr group, after Butler (1999). As Herring's (2004) discussion argues, quantitative techniques can be only a very partial view of a complex, somewhat subjective phenomenon, but are still a potentially useful exercise if understood with that qualification. Table 3 lists potential metrics with a description, an expanded explanation and statement about its availability through either the publicly accessible Flickr Application Programming Interface (API⁴) or data or Web-scraping (i.e., automatically extracting relevant data from the HTML source of relevant Flickr Web pages).

The set of metrics set out in Table 3 adds to Butler's (1999) work in a number of ways. Butler did not look at the length of existence of groups, but because some level of continuity is needed for community feelings to develop, it does seem a relevant measure. In relation to "Membership," it seemed useful to collect some data about the character of members, particularly given the claim of Flickr to be the "eyes on the world" (Naaman, 2006), implying an inclusive membership. We proposed looking at the degree to which members were in the United States time zone as a way of crudely judging how far Flickr is really a global system. We also wanted to look at the gender balance of membership. However, in practice, this information is not available. Although there is a field in the Flickr profile for a member's location, it is an uncontrolled text field, and so it would have been difficult to reliably establish the geographical location of people in the sample. Again, gender is not recorded in users' profiles, and so this could not be gathered.

In relation to the metrics proposed under "Communication activity," we wanted to look at the volume of activity. The number of discussion threads (10) relates specifically to forum activity, as opposed to photo-related activity. In relation to metrics of "Participation" structure, there were two aspects of this measured by Butler (1999), namely, the interactivity of postings and the level of participation of members. Items 11–16 are equivalents of the first category of metric; items 17–19 are equivalent to the second. Unfortunately all the possible interactivity measures were very hard to collect automatically. Because interactivity is likely to center on photos and photos in the pool can be surfaced in multiple places and reached in many ways, not just via a particular group, the extent of group interactivity is hard to measure. In addition, we added a new section, some measures of the formalization of the group, which relate to Herring's sixth aspect of virtual community.

⁴Publicly accessible Flickr API: <http://www.flickr.com/services/api/> (site accessed: 24 July 2010)

Methodology

To carry out the study, a two-stage methodology was used: (a) an initial empirical study of Flickr groups to provide overall group statistics based on a method to gather the numbers of groups; (b) a more detailed analysis of a random sample of 1,000 groups to derive characteristics of Flickr groups and validate the proposed metrics. Data were collected from Flickr using the publicly available API, combined with data-scraping for additional information not available through the API. The initial data was collected during September 2008 and the main sample in December 2008.

To analyze Flickr groups in-depth requires gathering a list of all possible groups from which to sample. A unique identifier references each Flickr group and this is required to gather further information about each group (either with the API or using data-scraping). However, Flickr provides no publicly available list of groups; the population is indeterminate. Our experience has shown that estimating the number of Flickr groups and creating a list of them is nontrivial. Initially we searched Flickr.com (group title and description) for the whitespace *<space>* character, assuming that most groups, regardless of language, would use such a character and would occur with high frequency. This approach returned a total of 461,272 groups (as of October 15, 2008). However, the Overall Flickr Group Statistics section shows that such results can be misleading because of effects, such as server load balancing. Results also only include public groups. To generate a list of group identifiers using this approach, we would have to execute multiple searches and extract the group identifiers from the HTML of the returned group pages. This is possible but requires multiple calls that, without careful consideration, can put a high load on the Web server host. Changes in Flickr.com now mean that searching using *<space>* returns no results and this approach to identify group IDs is not possible to use.

A more appropriate, and reproducible, approach is to use the publicly accessible API provided by Flickr. Many large-scale online services provide a publicly accessible API as a means of collecting data in a controlled and structured manner. However, similar to searching Flickr.com directly, the API also does not allow searches using characters, such as whitespace. Therefore, to compile a list of groups within the constraints of using the Flickr API, we used a form of dictionary lookup based on the following: (a) compiling lists of common words (or stopwords) in a range of languages⁵ (English, Spanish, Portuguese, Italian, French, German, Dutch, Japanese, Chinese, Romanian, Swedish, Polish, Finnish, Arabic, Russian, Czech, and Bulgarian), including punctuation markers ("approach 1"); and (b) using the 20 most popular Flickr tags (as of September 30, 2008) and translating these into multiple languages ("approach 2"). Translation was carried out automatically using the publicly

⁵Lists of common words in multiple languages are taken from the Snowball stemmer: <http://snowball.tartarus.org/> (site accessed: 24 July 2010)

TABLE 3. Theoretical Flickr group metrics and potential in gathering these metrics from Flickr.

Measure	Description	Explanation	Availability on Flickr groups
1. Length of existence	How long the group has existed	Some level of continuity seems to be a potential indication of the value of a group	Proxied by date of first upload to group pool
A. Membership			
2. Number of members		A crude measure of success of a group. Although having more members looks like more activity, very large groups are probably not functioning as communities	Available through API
3. Continuity of membership	Proportion of previous members who have stayed members in second time period	Arguably, the development of a group culture is only possible if there is continuity of membership, though a level of churn could also be seen as healthy	Date of joining not available
4. New members	Proportion of current membership that joined in last period (month/quarter)	Simpler variant on the above, equivalent to Butler's (1999) group growth	Date of joining not available
5. Internationalisation	Proportion of members from outside USA, either based on stated nationality or time zone of posting or time of day of posting	This would give us a sense of how diverse the groups were	Not available, because Flickr users do store location in their profile, this is an uncontrolled text field
6. Gender balance	Proportion of membership who are female	While topics may have different levels of interest between sexes, the balance of gender overall might be seen as an indicator of "health"	Not available, as not currently recorded in Flickr profiles
B. Communication activity			
7. Total number of photos posted			Available through API
8. Average number of photos per member	Total photos or recent photos divided by membership		Available through API
9. Number of recent Photos	Number of photos posted in last month/time period	A crude measure of activity	Date photo added to pool not available
10. Number of discussion threads	Total discussions threads ever initiated by group	Discussion about the group etc as opposed to simply uploading photos could be seen as evidence of group development	Available by screen scrape
C. Participation structure			
11. Levels of viewing of pooled photos		If photos in the group pool get a lot of hits, it suggests the group is active, although clearly one cannot simply attribute a high level of hits to group membership	Not available in the database

(Continued)

TABLE 3. (Continued)

Measure	Description	Explanation	Availability on Flickr groups
12. Number of recent photos, interactively posted	Number of photos posted in last month/time period and that are heavily clustered in a particular short time span	Implies a wave of activity in the group, where one set of postings sets off others	Hard to collect automatically
13. Level of in-group commenting	Proportion of all comments on a sample of group photos that are from group members		Hard to collect automatically
14. Uniqueness of group ("photo recycling")	Proportion of all photos that have only been posted to this group	If a photo is posted to lots of groups, presumably the photo is more important than the group. If the photo has only been added to this group it implies that the photo perfectly fits the purposes of this group and this group only.	Hard to collect automatically
15. Thread length (Discussions)	Average number of replies to discussion postings	The more replies to a discussion posting, the more evidence there is of group interaction	Hard to collect automatically
16. Message length (Discussions)	Number of words in messages	Longer messages implies more complex discourse	Hard to collect automatically
17. "Participation ratio" – Proportion of different members surfacing a photo in the group	Number of members who have recently (or ever) posted a picture as a proportion of total membership	Measures "rate of participation"	Not available
18. Core user dominance	Proportion of all photos uploaded by two most active members	Measures the extent to which the group is dominated by a few people	Available by screen scrape, as Flickr records number posted by most active and top five posters of photos
19. "Lurking"	Proportion of members who have not recently posted /ever posted an image		Very hard to collect automatically
D. Formalization			
20. Rules	Number of words in the rules of group	More complex rule making and description implies more activity in explaining the group. Unlikely to occur if the group is inactive	Available by screen scrape from group home page
21. Administration	Number of owners, administrators, moderators	The more effort is being put into organizing the group more formally, the more active and self-aware the group would seem likely to be	Available by screen scrape from membership listing, except where the group has chosen special names for Admin roles

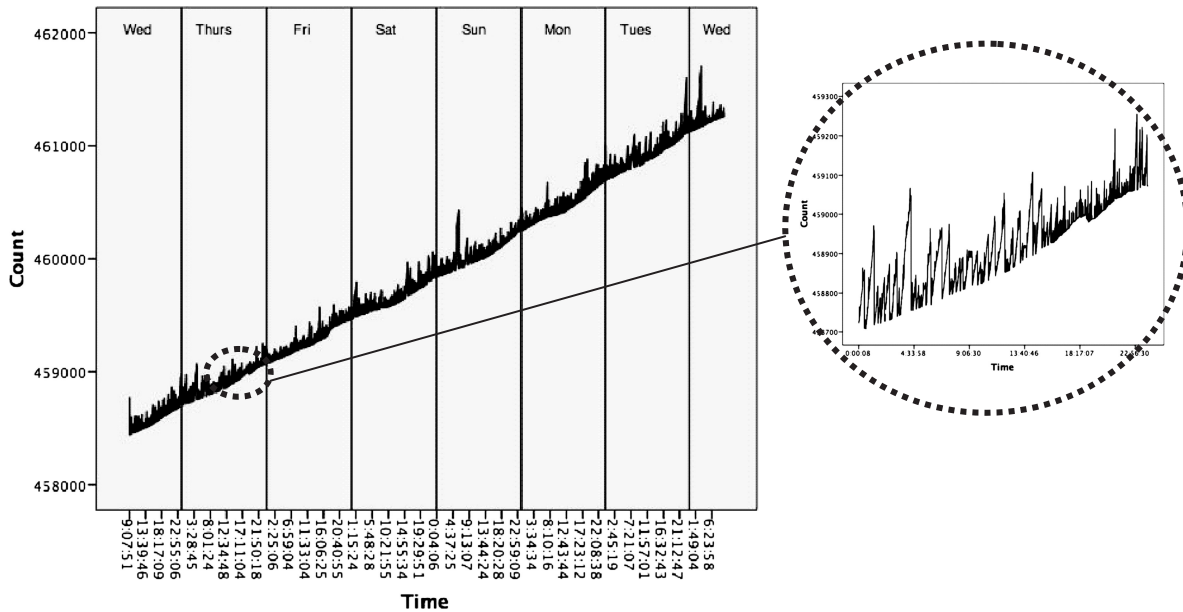


FIG. 1. Illustration of the rate of growth for the number of groups on Flickr (36,733 samples for 1 week period).

available Google Translate tools⁶ and verified manually. The Flickr API is then used to search for groups containing the given stopwords and tags. Using both approaches we were able to obtain a list (containing group identifier and name⁷) of 299,688 distinct groups in total (this is within the same order of magnitude as the previous estimate). The group identifier was used to gather more detailed information about each group using the Flickr API (e.g., “10017860@N00, Rock Climbing”). Although this excludes private groups, it gives an order of magnitude for Flickr groups and provides a large population from which to sample groups from and to gather data for the proposed group metrics.

From the list of total Flickr groups, we randomly sampled 1,000 groups and developed a customized crawler using a combination of the Flickr API and data-scraping. We were then able to gather detailed information about the groups based on our proposed metrics and enabled groups to be characterized and facilitated further analysis. Descriptive statistics for the sample group are as follows: number of members (mean = 1,054 members; standard deviation [SD] = 2,497) and the length of written description, in characters, about each group (mean = 1,314 members; SD = 2,646). To check the coverage of subjects for the sample group, a tag cloud was created for the top 200 tags from each group using the most frequent 100 tags. Comparing these tags to a tag cloud for all photos in Flickr on September 30, 2008 (the same time period as the sample), it was found that 84% of tags were the same,

thereby indicating a similar coverage of topics for the photos in the sample group compared with Flickr as a whole.

Information gathered for each group includes the following: (a) number of members; (b) description length; (c) first page of discussion history (most recent discussions); (d) number of uploaded photos; (e) number of discussions; (f) first and last uploaded photo; (g) 200 most popular tags (and font size); and (h) 5,000 most recently uploaded photos (if this was available). For each photo (approximately 2 million gathered from 1,000 groups), we gathered the following information: title, owner, tags, and date posted. For all owners from the photos collected previously, we also gathered information about the following: the owner’s geographical location, their Pro Account status, the total number of photos uploaded to date, and a number of attributes capturing social links (contacts, favorites, group membership). The information gathered for the 1,000 groups provides a wealth of detailed information with which to study group behaviors.

Results and Analysis

Overall Flickr Group Statistics

Estimating the total number of groups in Flickr is difficult because of constraints on searching Flickr.com and various system behaviors. For example, load balancing performed by the Web servers hosting Flickr causes unpredictable results, with the number of hits changing each time Flickr.com is queried (see Fig. 1). Also, the estimated number includes only publicly accessible groups. A more usable metric to capture the scale of Flickr is not an absolute figure for the number of groups but rather an estimate of the rate of change. To calculate this, we recorded the number of groups, plus uploaded photos and new users, that include white-space `<space>` in the title or description fields at 10s intervals

⁶Google Translate: <http://translate.google.com/> (site accessed: 24 July 2010)

⁷Example groups include: “10003140@N00, Widescreen Wallpaper”, “10005441@N00, ODD CHICAGO!!”, “10005981@N00, Tour Bermuda”, “10005982@N00, All About the Digit 5... Only.”, “10017860@N00, Rock Climbing”, “10007737@N00, Slow Children”

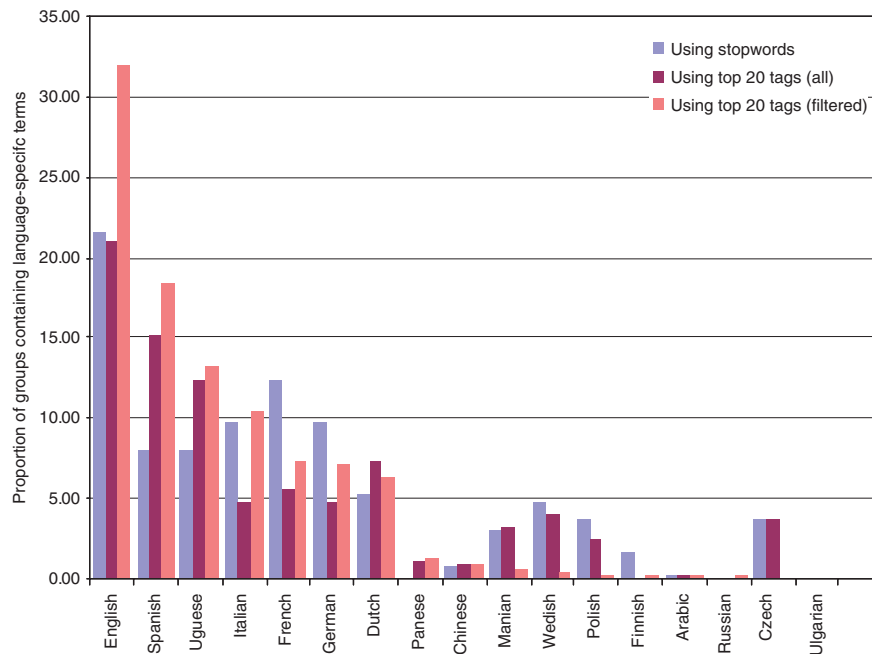


FIG. 2. Proportion of groups across languages (ordered by using filtered top 20 tags).

using Flickr.com during the course of 1 week (October 8 to 15 2008). This resulted in 36,733 data samples for (a) the number of groups, (b) the number of users, and (c) the number of photos. A steady increase in numbers should indicate new groups/user/photos being added, and over the course of 1 week, the effects, because of sampling, should be constrained.

The growth rates help to indicate the scale of Flickr and users interacting with the application. Figure 1 shows an example of the resulting growth rate for the number of groups on Flickr based on searching titles and descriptions with the whitespace character. The growth appears linear and fitting a linear regression line to the data gives the following: $y = 0.077x + 458447$ ($R^2 = 0.993$; $p < 0.001$). This suggests that, on average, a new group is added every 2.2 mins. Sampling the growth rate for the number of uploaded photos containing whitespace characters in the title or description, a total number of 1,011,909,272 publicly accessible photos have been uploaded. A resulting linear regression line of $y = 255.221x + 1.002 \cdot 10^9$ ($R^2 = 0.872$; $p < 0.001$) suggests that, on average, around 255 photos are uploaded in every 10 second interval, i.e., 26 new photos are uploaded per second. Finally, the number of registered Flickr users that are found using whitespace to search titles and descriptions is 30,480,789, with a resulting linear regression line: $y = 3.064x + 3.036e^7$ ($R^2 = 0.995$; $p < 0.001$). This suggests that, on average, a new member joins Flickr every 3.3 seconds. These results highlight the rate at which Flickr is changing and growing, a characteristic of most “mega Web sites.”

Using the data collection described in the methodology section, a total of 299,688 unique group identifiers (and titles) were gathered and form the population from which to sample groups. The aim of this stage was to gather a

comprehensive list of all group identifiers. Figure 2 shows the estimated language distribution (as a proportion of groups) of collected groups based on using (a) the lists of language-dependent stopwords, (b) the top 20 tags⁸ translated into multiple languages, and (c) a filtered version of the top 20 tags, whereby only tags that do not appear as tags in other languages are used. The last version gives a truer perspective on the language distribution because the same tags may be shared between languages (e.g., loan words), thereby mis-representing the language distribution of the group. Unsurprisingly, English and Spanish dominate the language of the groups, followed closely by other European languages.

Characteristics of the Flickr Group Sample

Based on data gathered for the proposed metrics in the Methodology section, we now present the results for each metric and discuss the implications for group behaviors. The results are categorized under membership, communication activity, communication structure, and formalization (see Results and Analysis subsections).

Membership

Length of existence. The length of existence of the group provides a possible indicator of whether there is likely to be a sense of community, on the crude assumption that elapsed time allows groups to become closer. Figure 3 shows the length of existence for the Flickr group sample. The length of existence is calculated using the date of the first upload

⁸Top 20 most popular tags (07/30/2008): water, white, light, portrait, flower, sunset, tree, yellow, girl, clouds, summer, fun, new, sea, photography, family, park, architecture, show.

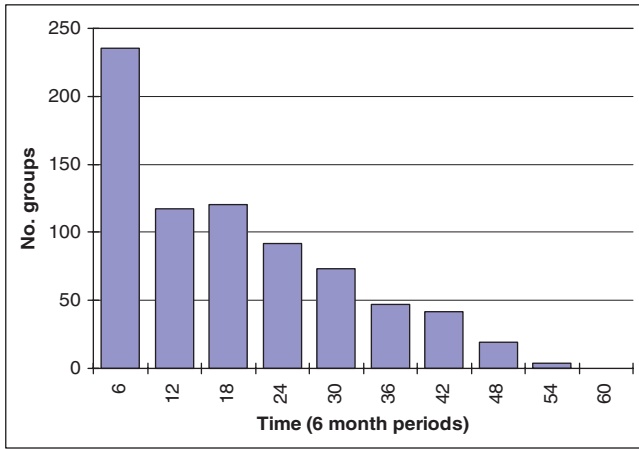


FIG. 3. Distribution of length of group existence (based on 6 month time intervals); $N = 749$.

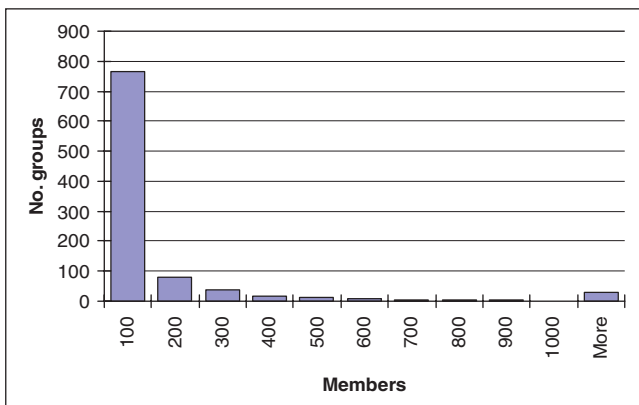


FIG. 4. Distribution of number of members for group sample; $N = 967$.

of a photo to the group photo pool as a proxy for the date of the foundation of the group. For 211 groups, no photo has been uploaded presumably because the group was created but never used, and so the date of creation cannot be inferred. For another 40 groups, unusable data was returned. For the 749 groups for which there was a date, the pattern indicates that nearly 50% of the groups had been created in the last year. It suggests either a surge of activity of group creation or that groups are commonly deleted⁹, for example, if they are not successful or conceivably they might be hidden from the searches through which the sample was collected.

Number of members. From the group sample, 20% had only one member, nearly 50% had less than 10, and nearly 80% of groups had less than 100 members (mean = 192; mode = 1; median = 12; max = 29,021); thus many groups are small and often akin to classic small groups examined in social psychology. Nearly 95% of groups have less than 500 members; only 3% (31 groups) have over 1,000 (see Figure 4). Figures regarding private groups cannot be publicly accessed, but these, by their very nature, would probably

⁹A group can only be deleted if it contains no members (i.e., when the last remaining admin person leaves the group).

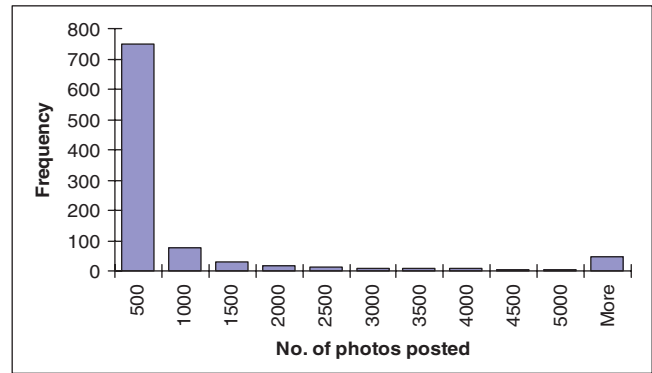


FIG. 5. Distribution of number of photos posted for group sample; $N = 967$.

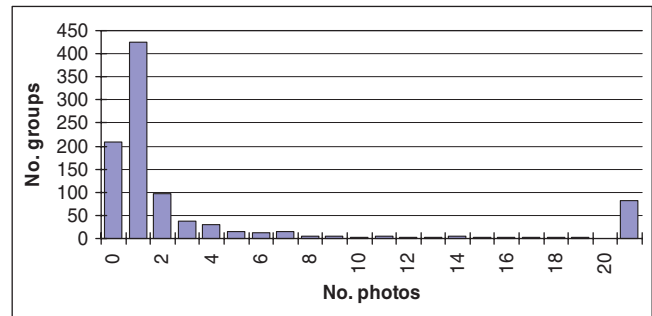


FIG. 6. Average photos uploaded per day; $N = 959$.

be small too, and so the impression of Flickr is that there are significant numbers of small groups. The very large groups are important and very visible, but there is an undergrowth of small groups that is easily overlooked. Butler's figure for the average membership of listservs was 163; the maximum was 2,245. The maximum for Flickr is vastly larger, showing how Flickr associations can be very different in quality, even if many groups might be operating in comparable ways to Yahoo! groups.

Communication Activity

Total photos posted. Twenty percent of groups had no photos, probably indicating groups that were created but never really used. Nearly 50% of groups had less than 100 photos (see Figure 5). A few groups have very large numbers of photos, with one group having over 1 million and three others over 100,000 photos (mean = 2,983; mode = 0; median = 79; max = 1,315,519).

Average number of photos uploaded per day. Forty-one groups with an existence of less than 1 day were excluded, because they produced absurd potential upload rates. In about 20% of groups, the average upload per day was 0 because no photos were ever posted. In 44% of all groups, the average was 1; in another 10% of groups, it was 2 (see Figure 6). This compares favorably for the activity levels on Butler's listservs, where the mean is only 1.635 messages per day (mean = 6.22; median = 0.58; maximum = 1,055).

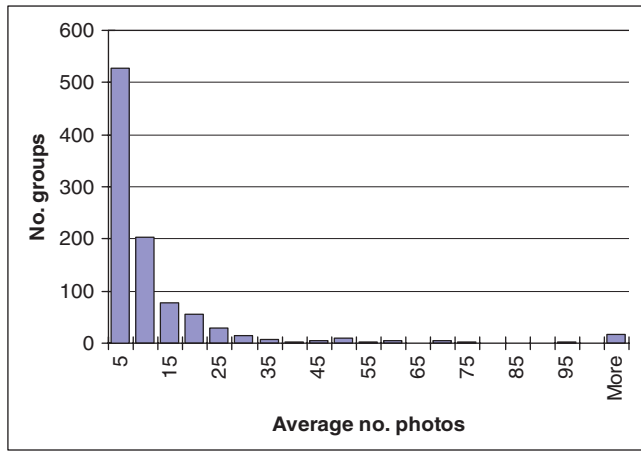


FIG. 7. Average number of photos in group pool per member; $N = 967$.

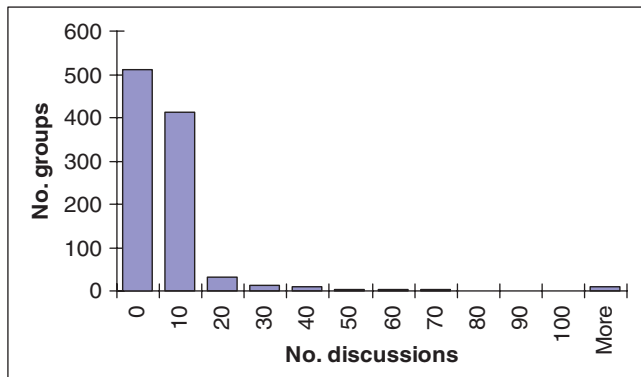


FIG. 8. Total number of discussions per group; $N = 1,000$.

Average number of photos per member. Over 50% of groups have less than five photos uploaded per member and around 80% have less than 10 photos per member (mean = 21.09; mode = 0; median = 4.41; maximum = 8,876; see Figure 7). These figures support Negoescu and Gatica-Perez's (2008) observation of low "group loyalty" (defined as a tendency to post most of one's photos to one group). There were 17 groups with more than 100 photos per member, only one had more than six members, which had 111 members. Thus, it does not seem to be the big groups that attract large numbers of contributions from a single individual. This is rather surprising as one would expect people to be attracted to putting photos where there is a large potential audience, i.e., a larger group.

Total number of discussions. Fifty percent of groups have never had any discussions; 40% of groups have had one discussion, i.e., one forum thread (see Figure 8). Nine groups had had more than 100 discussions. The figures support an interpretation that generally photos, with comments, not textual discussion, are the center of Flickr.

Communication Structure

Core user dominance. In around 20% of groups, the most active poster had posted only 10% of all photos; in another

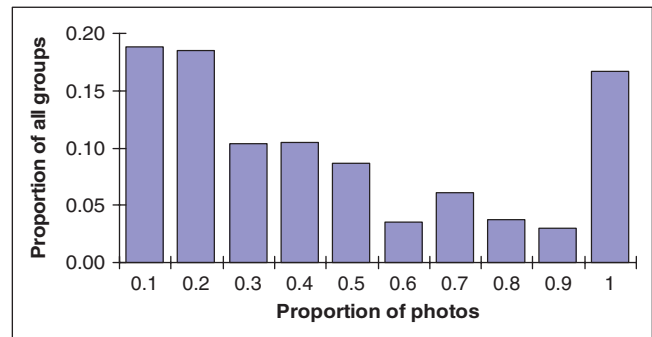


FIG. 9. Proportion of all photos posted by most active poster; $N = 769$.

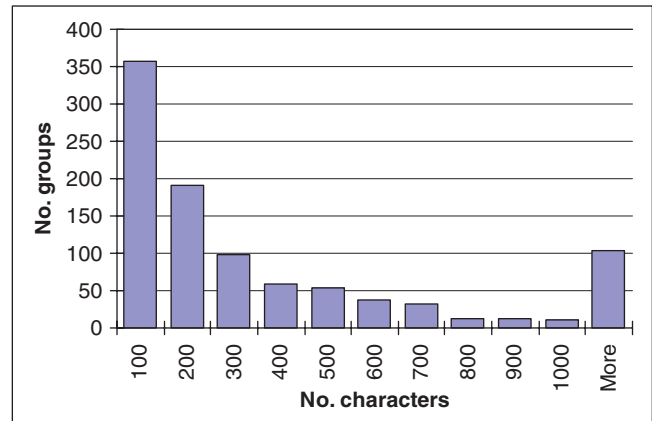


FIG. 10. Length of group description (in characters); $N = 967$.

20% of groups, they had posted more than 10% and less than 20% (see Figure 9). This, again, suggests significant numbers of groups where the power law of contribution does not apply; a few individuals are not dominating participation. In contrast, in another 17% of groups, one individual had posted between 90% and 100% of all photos; with 100 groups where 100% of photos were by one person. Yet these were all small groups with less than 10 members and around half of them had only one member. Butler (1999, p. 34) found that typically the two most active posters were posting more than 30% of postings. The figures are not easily comparable but imply lower levels of concentration, except in some small groups.

Formalization

Number of characters in rules of group. Sixty-six groups had 0 characters in their description and another 30% have less than 100 characters (see Figure 10), i.e., the description is no more than a sentence long (the sentence you have just read, up to the brackets, is 122 characters long). Nevertheless, there were over 100 groups with more 1,000 characters and two groups with more than 10,000 (mean = 484; median = 158; mode = 0; maximum = 23,581).

Group administration. Two hundred seventy-four groups in the sample had apparently no administrators (Admins). This could be because the admin had hidden their ID from

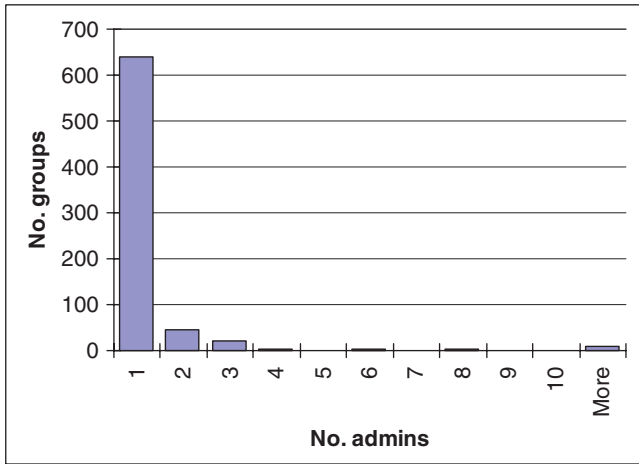


FIG. 11. Number of Admins; $N = 725$.

the listing or left a group after they had set it up. However, it was most likely because they had customized the name of the role, meaning that the approach used to gather the data would not collect information, as this was based on searching for the term “Admin” on the group Web page. This was unfortunate, as it would suggest that gathering data from the most active groups for this metric will be difficult. Nearly 90% of the groups with any admin have one only (see Figure 11). This is the default in Flickr, where the creator of the group is automatically admin. Ten groups had over 10 Admins (mean = 1.57; mode = 1; median = 1; maximum = 121).

To examine the pattern of moderation, the 25% of all the groups that had no Admins were not counted. Some groups do appear to have moderators without an admin, 90% of groups had no moderators, 18 had one (see Figure 12), and three groups had more than 10 moderators.

Characteristics of Large and Small Groups

Although the sample contained a lot of smaller groups, 20% of groups had more than 100 members. Typically, these had up to 1,000 members and a few (about 30) had more than 1,000. Such groups showed far greater activity: The average membership across them was 855 and the average number of photos posted was 13,721 photos. Not surprisingly, the profile of length of existence differed from the sample as a whole; nevertheless, 10% of the large groups had been created in the previous 6 months and a quarter was less than a year old. Interestingly, the top poster accounted for, on average, only about 3% of the photo pool in the large groups, and for around 90% of these groups, the top poster accounted for less than 20% of photos. Thus, there is not a sense of one or a few members dominating participation. Of the sample of 1,000 groups, 466 had 1–10 members and 202 had one member only (see Figure 13).

One hundred and ninety-four groups or 40% of groups with fewer than 10 members have no photos (see Figure 14); 149 of these only have one member. However, fifty have more than 100 and one group has 8,876 photos. It would require

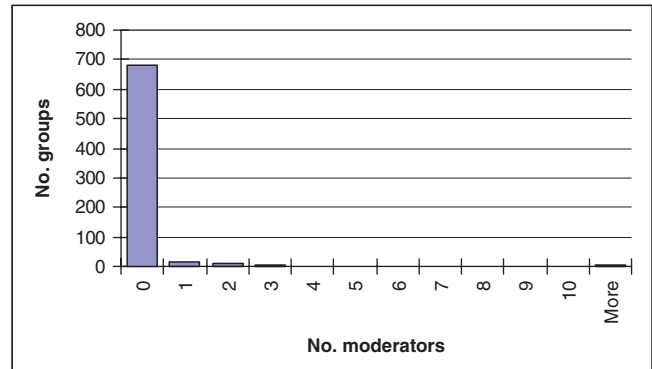


FIG. 12. Number of moderators, $N = 725$.

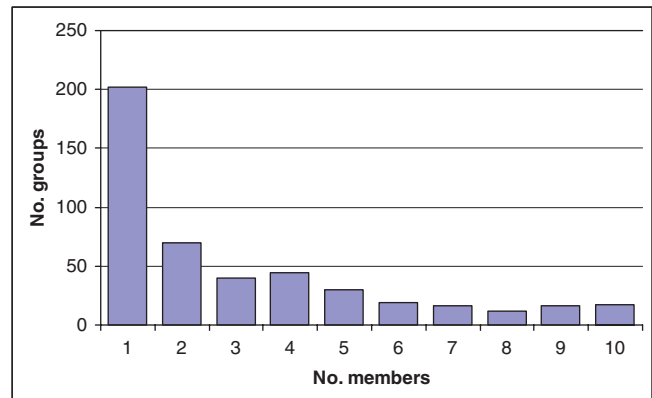


FIG. 13. Distribution of the number of members for groups with 10 or less members; $N = 466$.

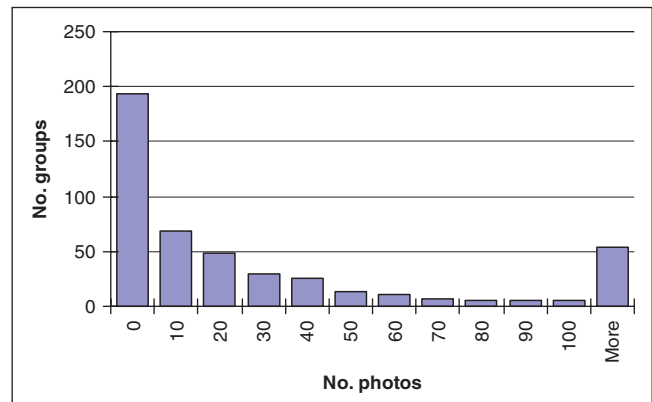


FIG. 14. Distribution of the number of photos for groups with 10 or less members; $N = 466$.

qualitative studies to confirm this, but the impression is that most small groups are not especially active.

Ten percent of such groups have no description (see Figure 15) and 37% have less than 100 character description. Thus, rather than being very active, the small groups are probably better understood as often being failed groups. Three quarters do not have much activity. This impression might be changed if we had data on private groups, which could be quite active with a restricted membership. Given the scale of Flickr, however, the fraction of small groups that are or

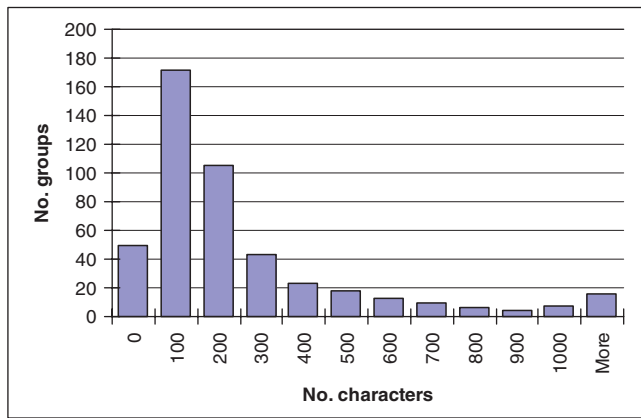


FIG. 15. Distribution of the length of descriptions for groups with 10 or less members; $N = 466$.

have been active provide an important part of the system as a whole. They could represent small scale sharing among friendship groups or project work.

Tables 4 and 5 provide details of the five smallest and largest groups from the sample (ordered by the number of members), indicating the diversity of groups exhibited within the sample.

Discussion

To produce a characterization of Flickr groups using automatically collected and publicly accessible data, we proposed emulating Butler's (1999) earlier study of Listservs to look at aspects of membership, communication activity, and communication structure in Flickr groups, using the API and data-scraping to collect data for a large sample of groups. Butler was interested in what the metrics for different online

groups suggested about the most appropriate metaphor to describe them, be that small group, voluntary association, or online collective. Taking into account the differences between photo sharing and online discussions, we proposed a number of other potentially interesting metrics. For example, we suggested looking at formalization of rules and roles. Although not all these metrics could be gathered automatically for Flickr, one contribution of the paper is to take forward a discussion about the most appropriate metrics for exploring the types of groups that emerge on large sites built from user contributions, one particularly adapted for analyzing Flickr. Applied to any particular group the metrics would be helpful in identifying those that might have a strong community character.

A second contribution of the article revolves around the accessibility of information about large sites such as Flickr. Despite the existence of much publicly available information for Flickr and an open API, there is much that is not or cannot be known about such sites. This arises from two main factors. First, some information is simply not recorded. Thus, although profiles on Flickr did originally have a field for gender, this was removed, and so it is not recorded in profiles. The place where the member is based is a free text field, and so the data recorded there is hard to process. It is also hard to interpret: The information could be about where the person originates or was born, but it could just be about where they are now. Second, the structure of the database makes it very hard to extract some forms of information through the API efficiently. What is available for groups is rather limited. Thus, it would be interesting to look at the extent to which group members commented on each other's photos. But the only way to measure this would be to scrape the IDs of (a sample of) photos in the group pool, then scrape commenter IDs from the photo's page, and then compare these to the list of

TABLE 4. Examples of the largest groups from the 1,000 group sample.

Group name	No. of members	Desc. length (chars)	Uploads by 1st poster	No. of photos	No. of moderators	No. of admins	Avg. upload/member
Flickrists: 29:010 of the BEST people on Flickr!!	29,021	4,002	9,018	1,315,519	0	0	45.33
DIGIFOTO Pro	11,239	4,749	399	170,181	0	8	15.14
Best of Cats (Invited photos only) - VOTE for your Pool Favorite	10,665	9,228	492	37,371	1	11	3.5
Fashion	7,828	12	1056	93,115	0	0	11.9
Nikon D50 Users	6,546	703	1365	163,755	0	1	25.02

TABLE 5. Examples of the smallest groups from the 1,000 group sample.

Group name	No. of members	Desc. length (chars)	Uploads by 1st poster	No. of photos	No. of discussions	No. of admins	Avg. upload/member
NitPickr	1	4,231	0	0	53	1	0
Bentley Gold Coast	1	1,894	0	0	1	1	0
Mallorca Radtour Frühling 2003 - Tour 1	1	1,203	3	3	0	1	3
Linn Drum Computer Owners Flickr Group	1	996	6	6	0	1	6
Red Bull CAN you Make it	1	949	0	0	0	0	0

group members (which would itself have to be scraped from the member listing). Thus, many of the questions one would like to ask are not easily answerable from the accessible data.

One small contribution of the article is to establish the order of magnitude of groups on Flickr. A sustainable approach is to use a list of common stopwords and tags, translated in multiple languages. The approach yielded an estimate of around 300,000 public groups in September 2008. However, the difficulties of the calculation further underline the inaccessibility of data about Flickr.

The fourth main contribution is what we learn from the metric data that could be gathered. It is difficult to reach very firm conclusions without some point of comparison. Butler's (1999) data may simply be from too different a context to offer such a reference point. What stands out, however, is the large number of groups with only a handful of members and low levels of activity. Nearly 50% of groups have less than 10 members; 20% have only one. The undergrowth of small groups may be an important part of the ecology of Flickr, supporting fleeting project work or the activity of small friendship group, an outgrowth of ic network activity. On the other hand, there are also some extremely large groups, whose scale is very much greater than those seen in Butler's work, in terms of numbers of members and contributions. These could be rather impersonal places where a high level of activity prevents individual voices being heard. The big groups tend not to be dominated by a few individuals. So, although many writers have talked about a power law in participation, this does not seem to apply to Flickr groups. Core user dominance is much more a feature of small groups and perhaps a sign of "failure." Between the two extremes of very large and very small groups are many groups that support membership sizes and patterns of activity that are consistent with groups with a sense of community.

Web2.0 as the "social web" is different in quality from the world of online communities as sustained on Usenet, listservs, or forums. The social network model of design seems to be superseding, or more accurately overlaid, on the online community model. Yahoo! groups, for example, organize social activity around participation and interaction in interest groups, where the group is a central reference point. In the typical social network model, the individual is central, and membership of groups more fleeting and less exclusive (boyd 2006). Baym suggests that even in social networking contexts people with similar interests tend to encounter each other again and again with the result that community emerges but is distributed "throughout a variety of sites in a quasi-coherent network pattern" (Baym 2007). Flickr is slightly different in that the whole social network is itself already dedicated to a common broad interest, photography. The chances of encountering individuals with very similar tastes and interests are increased and group formation can happen within Flickr itself, rather than distributed across a number of places. Much Flickr activity is not centred on groups. Fifty percent of people do not ever post to groups (Negoescu & Gatica Perez, 2008). Many individuals are focussed on their own photos. The site is social in that content is created by

a mass of individuals, but the levels of direct interaction are not always very high. Yet alongside the ego-centred activities, groups do form and the large number of Admins and moderators point to a level of emerging online community feeling too.

Conclusions and Future Work

This article has explored the role of groups on Flickr. It has discussed suitable metrics for examining aspects of membership and activity levels, especially to capture some overview of the types of community that are supported on Flickr. These metrics could be applied more generally than Flickr to capture characteristics of various online groups or communities. Indeed, the value of such metrics would be greatly enhanced given comparative data. It would be interesting to see how the group structures differ on a site like YouTube. It has shown how many Flickr groups there are and has also begun to show what Flickr groups are typically like. It has pointed to the importance of small groups on Flickr. This was not apparent before because the large active groups are more visible, e.g., in the search results. Thus, although it is vast in terms of numbers of photos and members, groups in Flickr operate at a more human scale. This might be one aspect of its success: that nested within the larger groups there are patches of activity and organization. Some feel for the overall character of groups on Flickr allows us to benchmark specific groups. The article has also demonstrated some quite profound obstacles to fully exploring the character of participation, at least to doing so efficiently.

It is not part of our suggestion that groups are necessarily central to the experience of Flickr for many users. Only committed users join groups. But it is one aspect of Flickr, and clearly in complex ways it has contributed to the success of the design. Understanding how individual-centred activity such as organizing one's own photos or commenting on those of others links to online community activity in such Web sites is important. Qualitative studies could examine the character of different-sized groups, looking more closely at the different types of groups, how these relate to different photographic practices, as well as motivations to participate and group feeling.

Acknowledgments

This research was partially funded by the EU Marie Curie ToK Grant Memoir (MTKD-CT-2005-030008) and FP7 integration project LivingKnowledge (Contract No. 231126). We thank the three anonymous reviewers for their thoughtful and helpful suggestions.

References

- Angus, E., Thelwall, M., & Stuart, D. (2008). General patterns of tag usage among university groups in Flickr. *Online Information Review*, 32(1), 89–101.
- 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 (pp. 44–54). New York: ACM Press.
- Blanchard, A. (2007). Developing a sense of virtual community measure. *Cyberpsychology and Behavior*, 10(6), 827–830.
- Blanchard, A., & Markus, M.L. (2004). The experienced “sense” of a virtual community: Characteristics and processes. *Database for Advances in Information Systems*, 35(1), 65–79.
- Boyd, D. (2006). Friends, friendsters, and top 8: Writing community into being on social network sites. *First Monday*, 11(12). Retrieved from <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/rt/printerFriendly/1418/1336>
- Butler, B.S. (1999). When is a group not a group: An empirical examination of metaphors for online social structure. Katz Graduate School of Business, University of Pittsburgh. Retrieved from <http://www.pitt.edu/~bbutler/KatzWorkingPaper2001.pdf>
- Castro, M.C. (2006). The Macuarium set of community of practice measurements. Retrieved from http://www.macuarium.com/cms/index.php?option=com_remository&Itemid=0&func=fileinfo&id=447
- Chin, A., & Chigwell, M. (2006, August). A social hypertext model for finding community in blogs. Paper presented at Hypertext 06, Odense Denmark.
- Cothrel, J.P. (2000). Measuring the success of an online community. *Strategy and Leadership*, 28(2), 17–21.
- Cox, A.M., Clough, P., & Marlow, J. (2008). Flickr: A first look at user behavior in the context of photography as serious leisure. *Information Research*, 13(1). Retrieved from <http://informationr.net/ir/13-1/paper336.html>.
- Cox, A.M. (2008). Flickr: A case study of Web2.0. *Aslib Proceedings*, 60(5), 493–516.
- Cummings, J.N., Butler, B., & Kraut, R. (2002). The quality of online social relationships. *Communications of the ACM*, 45(7), 103–108.
- Davies, J. (2006). Affinities and beyond! Developing ways of seeing in online spaces. *E-learning*, 3(2), 217–234. Retrieved from http://www.words.co.uk/elea/content/pdfs/3/issue3_2.asp#8
- Herring, S.C. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In S. Barab, R. Kling, & J.H. Gray (Eds.), *Designing for virtual communities in the service of learning*. New York: Cambridge University Press.
- Malinen, S. (2010). Photo exhibition or online community? Proceedings of the Fifth International Conference on Internet and Web Applications and Services (pp. 380–384).
- McMillan, D.W., & Chavis, D.M. (1986). Sense of community: A definition and theory. *Journal of Community Psychology*, 14, 6–23.
- Miller, A.D., & Edwards, W.K. (2007). Give and take: A study of consumer photo-sharing culture and practice. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07) (pp. 347–356).
- Naaman, M. (2006). Eyes on the world. *IEEE Computer*, 39(10), 108–111.
- Negoescu, R.A., & Gatica-Perez, D. (2008). Analyzing Flickr groups. Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval (CIVR '08) (pp. 417–426).
- Nov, O., Naaman, M., & Ye, C. (2008). What drives content tagging: The case of photos on Flickr? In Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems (CHI '08) (pp. 1097–1100). New York: ACM Press.
- Nov, O., Naaman, M., & Ye, C. (2009). Analysis of participation in an online photo-sharing community: A multidimensional perspective. *Journal of the American Society for Information Science and Technology*, 61(3), 555–566.
- Perez, J.C. (2007). Flickr co-founder talks about growth plans. Retrieved from <http://www.macworld.com/article/56711/2007/03/flickr.html>
- Prieur, C., Cardon, D., Beuscart, J.-S., Pissard, N., & Pons, P. (2008). The strength of weak cooperation: A case study on Flickr. Retrieved from <http://arxiv.org/abs/0802.2317>
- Recuero, R. (2007). Communities in social networks: A case study of Brazilian fotolog. In Proceedings of Internet Research 8.0: Let's Play, 2007 (AIR 8.0).
- Rorissa, A. (2010). A comparative study of Flickr tags and index terms in a general image collection. *Journal of the American Society for Information Science and Technology*, 61(11), 2230–2242.
- Schoberth, T., Preece, J., & Heinzl, A. (2003). Online communities: A longitudinal analysis of communication activities. In Proceedings of the 36th Annual Hawaii International Conference on System Sciences (Vol. 7, p. 216a). Washington, DC: IEEE.
- Sigurbjörnsson, B., & van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In Proceedings of the 17th international Conference on World Wide Web (WWW '08) (pp. 327–336).
- Smith, M. (2002). Tools for navigating large social cyberspaces. *Communications of the ACM*, 45, 51–54.
- Stivilia, B., & Jörgensen, C. (2009). User-generated collection-level metadata in an online photo-sharing system. *Library & Information Science Research*, 31(1), 54–65.
- van House, N., Davis, M., Ames, M., Finn, M., & Viswanathan, V. (2005). The uses of personal networked digital imaging: An empirical study of cameraphone photos and sharing. In Proceedings of the 23rd Annual SIGCHI Conference on Human Factors in Computing Systems (CHI '05) (pp. 1853–1856). New York: ACM Press.
- Van Laere, O., Schockaert, S., & Dhoedt, B. (2010). Towards automated georeferencing of Flickr photos. Proceedings of the Sixth Workshop on Geographic information Retrieval (GIR '10) (pp. 1–7). New York: ACM Press.
- van Zwol, R. (2007). Flickr: Who is looking? In Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (pp. 184–190). Washington, DC: IEEE.