Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

How Content, Community, and Engagement Affect the Life Cycles of Online Photo Sharing Groups

Lyndon Kennedy
Yahoo Labs
lyndonk@acm.org

David A. Shamma
Yahoo Labs
aymans@acm.org

Abstract

We investigate the life cycles of groups on online photo sharing sites to determine the factors that contribute to their growth. We propose features to capture the social, engagement, and content characteristics of groups and conduct a large-scale retrospective analysis of more than 170k groups on Flickr to discover the relative importance of each of the factors on the long-term growth of groups. This yields models that can predict the success/death of groups with upwards of 82% accuracy.

1 Introduction

Online social networks can serve a number of purposes, ranging from maintaining personal or professional communication networks (e.g. Facebook, LinkedIn), following updates from interesting sources (e.g. Twitter, Tumblr), or engaging around shared media (e.g. Flickr, YouTube). In this work, we investigate online groups on media-centric social sharing sites. We hypothesize that a number of factors, such as group popularity, engagement, reflection of social ties, and cohesiveness of topical content might all affect the long-term viability of any given group. To measure these effects, we design a series of features to capture the various aspects of groups that can affect their longevity.

We conduct experiments using a large-scale data collection from Flickr. We track the groups that users form on Flickr and retrospectively recreate their historical states over extended periods of growth. A variety of engagement measurements, social network analyses, and content analysis techniques are used to measure the engagement, social, and topical characteristics of groups over time. We use statistical methods to analyze the relative importance of each of the designed features and to build models capable of predicting group growth from early activity observations.

We find that from observing a group only one month after its creation, we can predict the long-term growth of the group with 82% accuracy. We can further characterize the various contributing factors to growth. The users' perceptions of increased engagement, the presence of a unified topic with diverse visual perspectives, and strong social interconnections all play a positive roles in the continued growth of a group. Indicators show that users sometimes create groups for single-use purposes, with no intention of the group existing in the long term. Additional findings suggest that vibrant groups also exist around the meta-subject matter of photography, itself.

2 Background and Related Work

Flickr is an online photo sharing platform where users can create online identities and upload and share their photos. Once uploaded, the photos can be further enriched by additional metadata such as brief textual descriptions, geolocations, tags, and labels of people within the photos. Users can create unidirectional social connections with each other, similar to the "follow" action on other platforms. Flickr also features "groups," which are informal communities where users can participate in discussion boards or contribute photos to the group "pool."

The spread of social media sharing sites like Flickr has made the study of social networks increasingly tractable from a computer science point of view (Backstrom et al. 2006; Kairam, Wang, and Leskovec 2012) and some work has been dedicated to understanding interactions in Flickr groups (Nov, Naaman, and Ye 2008). De Choudhury (2009) investigated the prediction of group activity in Flickr groups using social actions between group members and Cox, Clough, and Siersdorfer (2011) characterized Flickr groups based on membership, communication activity, and structure. Other work (Negoescu et al. 2009) has shown how Flickr groups might be meta-organized by their distributions of tags into latent topics, while Prieur et al. (2008) show the relationships between topicality and social connectivity in groups. Other lines of research have used groups as a starting point for learning computer vision models for visual concept detection Ulges, Worring, and Breuel (2011). Our work extends past these prior works by taking a multimedia perspective to characterizing group structures based on a variety of cues. We then further develop a framework for considering all of these diverse cues and examining the implications that they have on whether a group will continue to grow over an extended time frame.

3 Data

We obtained a dataset of public groups on Flickr. The set consists of the group names, descriptions, and creation timestamps. This also includes all group members, all photos added to the group, and the respective timestamps of members joining and photos being added. We also gather timestamped social connections between users and additional metadata about the photos, such as their tags, the images, and the activity (views, comments, favorites) that they have received. From this dataset, we can retrospectively construct views of groups and how they were composed of people and media at any historical point in time by applying thresholds on timestamps for members joining and photos being added with respect to the time of groups creation. We limit our analyses to the public groups that have accumulated more that 50 photos in their pool in the first month since the founding of the group. This yields more than 174k groups spanning the past 12 years (since Flickr's launch).

4 Predictor Features

4.1 Prior Growth

Groups that have more members and more photos might continue to grow in the future. To measure this, we simply take the increases in the number of members of the group and the number of photos in the pool over the course of some previous time period (for example, the past month).

4.2 Engagement Lift Features

A common reason that users participate in social media sharing sites is to gain exposure and receive engagement from other users. To measure this, we apply a method for quantifying the rise in engagement that a user expects from submitting a photo to a group. For each user, we examine his/her public collection of photos and calculate the number of views, favorites, and comments that an average photo would receive. For each photo in a given group, we calculate the ratio of the true number of views, favorites, or comments that the photo has received to the expected values of these engagements for the owner of that photo. We aggregate these ratios over each photo in the group to measure the expected lift in engagement that photos in the group have compared to generic photos from each user.

4.3 People and Identity Features

Factors related to the identity of group participants and their relationships to each other might also affect group growth.

Contact Density We consider the structure of social connections between members of each group. We count the number of social connections between group members and take its ratio against this maximum value of connections in a fully-connected graph composed of group members.

Owner Diversity To measure the diversity of group contributors, we calculate a multinomial distribution over the owners of each of the photos that have been contributed to the group's pool as the probability of a photo coming from owner $P(O_i)$ is the ratio of the number of photos contributed by owner i, $\#(O_i)$, to the total number of photos in the pool $|G|$, or $P(O_i) = \frac{\#(O_i)}{|G|}$. We then characterize this distribution by calculating the ownership entropy, $H(O)$ using: $H(O) = -\sum_i P(O_i)\log(P(O_i))$.

4.4 Photo and Theme Coherence Features

The degree to which photos in a group are related to a single visual topic might also reflect the lifespan of the group.

Visual Diversity We represent the content of the images in by applying computer vision (Krizhevsky, Sutskever, and Hinton 2012) to create classifiers for common concepts (such as "cat," "wedding," "graffiti," etc.), yielding models for more than 1500 visual concepts. Each image in each group in our dataset is then passed through this pipeline, giving us a set of automatic "visual concepts." We calculate a multinomial distribution of these visual concepts in the group as $P(V_i) = \frac{\#(V_i)}{|V_G|}$, where the probability of a photo having visual concept i is estimated as that ratio of all photos having visual concept i in the group pool, $\#(V_i)$, to the total number of visual concepts detected over all photos in the group, $|V_G|$. We calculate the entropy of this distribution, as above, to characterize the visual diversity of the group.

Tag Diversity Flickr provides users with an easy way for users to manually tag their photos. We calculate a multinomial distribution of tags in a group as $P(T_i) = \frac{\#(T_i)}{|T_G|}$, where the probability of a photo having tag i is the ratio of all photos having tag i, $\#(T_i)$, to the total number of tags applied to photos in the pool, $|T_G|$. We characterize this distribution, again, by calculating the entropy.

4.5 Group Descriptive Terms

When users create groups on Flickr, they provide names and descriptions for the group. We extract features related to the terms used in these descriptions by some lightweight text analysis. We tokenize the strings into unigrams and then further normalize them by stripping out punctuation and converting to lowercase letters. We limit our analysis to the approximately 4000 terms that have appeared in the names of at least 100 of the 174k groups in our set.

5 Statistical Modeling

5.1 People, Topics, and Attention

To understand the relative predictive power of the features that we have designed, we apply a logistic regression. We specify the dependent variable as whether the group is growing in terms of photos added on a month-by-month basis *faster* or *slower* than the median after a period of *one year*. The independent variables are defined by the above-described measures of engagement, social connectivity, and subject diversity. These independent variables, however, are measured at the point in time when the group is only *one month* old. The variables are all normalized (zero mean, unit variance) such that the resulting coefficients are directly comparable. Effectively, we are taking a short-term view of a new group and seeing which measurable aspects of the group will be predictive of its long-term staying power.

5.2 Descriptive Terms

Analyzing the relative predictive power of terms requires the ability to incorporate many sparse predictors that may be highly correlated. To achieve this, we adopt a penalized

Category	Feature	β	p
Intercept	Intercept	-4.293	***
Prior Growth	Number of Photos	0.000	***
	Number of Members	0.001	***
Engagement	Favorites	-0.008	*
	Views	0.631	***
	Comments	0.009	**
People	Contact Density	0.562	***
	Owner Entropy	0.498	***
Photos	Visual Diversity	0.360	***
	Tag Diversity	0.470	***

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

Table 1: Coefficients of logistic regression for group growth above the median after 1 year.

p1, etsy, hdr, comment, photographers, colors, bw, fujifilm, admin, hand, olympus, canon, toys, postcards, panasonic, guess, 365, nikon, cats, blythe, macro, 52, pentax, ♥, award, users, please, leica, solo, yellow, trees, vintage, doll, artists, moments, il, landscapes, lovely, japanese, rules, post, aircraft, fuji, handmade, beautiful, dolls, flickrs, children, today, portraits, etc, cars, trucks, animal, dogs, disney, uk, ford, españa, british, flowers, daily, anything, flickr, photographer, graffiti, perfect, lovers, mundo, views, arte, trains, dark, pretty, passion, now, faces, magazine, models, sony, random, photographs, videos, wildlife, e, architecture, parks, birds, toy, creations, lego, nature, buses, books, cat, fotografia, bird, people, france, cute

Table 2: The 100 strongest terms for predicting that that a Flickr group will grow *above* median rate in 1 year.

logistic regression approach (Friedman, Hastie, and Tibshirani 2010), which can predict a binary dependent variable while guarding against collinearity of independent variables. Again, we set the dependent variable to be group growth rate at one year and use input variables after one month of group existence. Similar to Gilbert (2012), we code the terms as either present (1) or not present (0) in the description of each group and learn the model using this joint feature space.

6 Results and Discussion

Table 1 shows the coefficients of the logistic regression. We see that most factors are positively related to long-term growth. Together, they predict growth with 82% accuracy.

6.1 Predictive Features

We see that the propensity for a group to contribute positively to the overall engagement that a photograph receives is somewhat positively predictive of the long-term growth of the group. This observation supports the earlier hypothesis that exposure and engagement are important motivators for any action taken on social media sharing sites.

The composition of the people that make up a group and the distribution of their participation rates are both predictors of group longevity. The contact density of the social graph of group participants is a strong predictor, which indicates that groups where users have established direct social connections outside of the group have a higher long term success rate. The degree of diversity in users submitting photos is also a strong predictor, indicating that stronger groups are formed by diverse participants submitting photographs.

The topical diversity of the photos contributed to a group pool, both in terms of the visual content of the photos and the tags associated with them are predictive of long-term group growth. Interestingly, in both cases, the *more diverse* a group is in terms of its topical or visual content, the more likely it is to have sustained long-term interest. This suggests that even in the case of groups formed around single subject matters, varied viewpoints are valued over repetitions of the same perspectives.

6.2 Topics for Growth

We show the most predictive terms for long-term group growth in Table 2 and discuss observed patterns below.

Participation. An interesting effect that we observe is the appearance of terms around the expected engagement practices. One of the top terms is "comment," which reflects that some groups have rules where users posting photos are expected to add comments or favorites to other users' photos in order to incentivize participation ("please," "post," "views," and "rules" are also predictive terms related to this practice). Another participatory practice is giving awards to photographs (the "award" term appears in the list as does the "♥" glyph). Finally, we also note the appearance of terms related to long-term group participation projects, such as "365," where users post a new photo on each day of the year, or "52," where new photos are posted every week. By their nature these types of interactions predict sustained contributions over time periods much longer than just the first few days or months of the group. Interestingly, this is at odds with the above-described observation that increased in engagement due to groups are not strongly predictive of group longevity. This may be due to the fact that these types of rule-based groups might only increase the users' *perception* that the group is increasing engagement when it is not significantly increased in practice.

Subject Matter as Interest. Many of the terms that appear to be predictive of group survival are related to shared common interests and photographic subject matters. Some examples of these are natural landscapes ("landscape," "flowers," "wildife"), vehicles ("buses," "trains," "cars"), and people ("portraits," "models"). All of these suggest that groups about single topics of interest are likely to grow. This is at odds with our above-described finding that visual and topical diversity are also predictive of success. This might indicate that the types of photos in these groups are related to similar topic matters, but show a diverse set of viewpoints and perspectives on the subject.

anh, graduation, seo, 07, 2015, wedding, beckles, reunion, birthday, 2012, 2014, 2013, 09, 2011, trip, 2007, 2010, 2008, 2006, 08, 2009, 2005, grade, anniversary, jonas, workshop, website, camp, office, chris, clan, course, annual, scott, chapter, vs, curso, conference, program, family, photowalk, weekend, marketing, ve, party, event, student, summit, march, mission, foundation, rica, vacation, canyon, fall, study, tour, david, period, training, soccer, basketball, run, institute, class, team, walk, youth, 12, media, tech, ball, swap, school, adventure, memorial, celebration, august, june, exhibition, internet, fc, ski, page, gang, competition, july, 13, portfolio, bar, adventures, alumni, band, paul, ii, resort, hockey, site, baseball, college

Table 3: The 100 strongest terms for predicting that that a Flickr group will grow *below* median rate in 1 year.

Photography Practice as Interest. The high prevalence of terms like "photographers," "photographs," and "fotografia" among the terms predictive of group growth is surprising since all groups are fundamentally grounded around the sharing of photographic content. However, the fact that these groups specifically reference the practice of photography is indicative of the importance of groups that cater to the interests of photography hobbyists. This implies that an important function of groups on a media-centric sharing site like Flickr is for users who do not necessarily know each other to connect socially around their shared enthusiasm for the practice of photography.

6.3 Purposefully Short-Term Groups

In the results shown in Table 3, we see the terms that are most negatively associated with long-term activity in groups. We see overwhelmingly that these terms are temporal in nature, referring to specific events or points in time. Terms corresponding to years ("2015," "2014," "09," etc.) as well as months and seasons ("july," "august," "fall") are among the strongest negative terms. Terms related to common one-time events ("graduation," "wedding," "reunion,") are also strong as are sporting events ("basketball," "soccer," "hockey") and terms related to vacations ("vacation," "canyon," "resort," "ski"). These effects are due to the practice of using Flickr to create ad hoc groups for the purposes of sharing photos after an event between a number of users who may have been in attendance, without any intention of creating a long-term interest group. This might indicate the need for designing tools that are more tuned to this specific use-case.

7 Conclusions

We have presented a large-scale investigation of social groups on media-centric social networking services. We designed several methods for capturing the structure, composition, and engagement effects present in a group and implemented statistical analyses to understand and leverage the predictive powers of each of these cues. This resulted in a system capable of predicting group longevity with more than 82% accuracy. The analyses further yielded insights into the relative importance of social connectivity within groups and diverse contributions and viewpoints around topics of interest. Our investigation further uncovered indications of other uses of groups, such as single-purpose groups for disseminating media between event attendees and the staying power of groups formed around the media-creation process, itself, rather than any other sort of social or topical connectivity.

References

Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; and Lan, X. 2006. Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, 44–54. New York, NY, USA: ACM.

Cox, A.; Clough, P.; and Siersdorfer, S. 2011. Developing metrics to characterize flickr groups. *Journal of the American Society for Information Science and Technology* 62(3):493–506.

De Choudhury, M. 2009. Modeling and predicting group activity over time in online social media. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, HT '09, 349–350. New York, NY, USA: ACM.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 33(1):1.

Gilbert, E. 2012. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, 1037–1046. New York, NY, USA: ACM.

Kairam, S. R.; Wang, D. J.; and Leskovec, J. 2012. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, 673–682. New York, NY, USA: ACM.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.

Negoescu, R.-A.; Adams, B.; Phung, D.; Venkatesh, S.; and Gatica-Perez, D. 2009. Flickr hypergroups. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, 813–816. New York, NY, USA: ACM.

Nov, O.; Naaman, M.; and Ye, C. 2008. What drives content tagging: The case of photos on flickr. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, 1097–1100. New York, NY, USA: ACM.

Prieur, C.; Pissard, N.; Beuscart, J.-S.; Cardon, D.; and Pons, P. 2008. Thematic and social indicators for flickr groups. In *ICWSM*.

Ulges, A.; Worring, M.; and Breuel, T. 2011. Learning visual contexts for image annotation from flickr groups. *Multimedia, IEEE Transactions on* 13(2):330–341.