Assessing the Impact of the Social Network on Marking Photos as Favorites in Flickr

Luiz Alves, Matheus Maciel, Lesandro Ponciano, Andrey Brito
Departamento de Sistemas e Computação
Universidade Federal de Campinta Grande
Campina Grande, PB, Brazil
{luiz.filho,matheus.maciel}@ccc.ufcg.edu.br {lesandrop, andrey}@lsd.ufcg.edu.br

ABSTRACT

Understanding user behavior in user generated-contend systems (UGCs) is a fundamental building block for maximizing usefulness of these systems. For example, the importance of users' connections and the degree of iteration is needed in order to build meaningful recommendation systems. In this paper, we evaluate the impact of users' network on marking photos as favorites in Flickr considering various levels of distance between users in the network topology. We have observed that users' first-degree contacts are clearly the biggest contributors to the evaluation of hers/his photos (i.e., marking them as favorites). This result indicates that, in contrast to other UGCs such as Twitter and Digg, contact relations are a good indication of content interests in Flickr.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Miscellaneous—*Information Systems Applications*

Keywords

Social Network Analysis, User Influence, Flickr

1. INTRODUCTION

In recent years there has been great expansion of user-generated content systems (UGCs) such as Twitter, YouTube, Flickr and Digg. These systems are used daily by millions of users with the goal of sharing and/or consume content from various subjects such as art, politics and economics [6, 7, 9].

Several studies have investigated how content shared in a UGC is valued by users and disseminated through their contacts in the social network [3, 4, 8, 13]. These studies have focused, for example, on characterizing how content is propagated within the network and on identifying the most popular and the most influential users [3]. Some studies have shown that users maintain large networks of contacts, but these networks have little effect on the content they consume [8]. This challenges the belief that relationship between two users indicates common interests in the types of published content.

The present work analyzes the impact of users' network on marking contents as favorites in the social network, considering various levels of distance between users in the network topology. We examine the impact of the users' network on marking photos as favorites in Flickr[1]. Flickr is a popular photo-sharing site with millions of users, more than 6 billion[2] photos posted, and thousands of new photos being posted every day. In Flickr, users can perform various activities such as: creating contacts between one another, posting photos, and marking photos as favorites. Our analysis takes into account the contact network formed by contact links among Flickr users. We analyze different distance levels between users in order to identify the relationships between the users' preference for the content posted and the users' proximity in the social network topology.

Our results show that 70% of the photos marked as favorites by users are posted by their contacts. This indicates that users show interests for contents posted by other users with whom they are connected in the social network. Moreover, unlike the behavior observed in the systems Digg.com [8] and Twitter [3], in Flickr, contacts' network has potential as a way of disseminating content. Our results show that approximately 10% of the photos marked as favorite by users have not been posted by their contacts, but disseminated by them via favorites marks. In general, the behaviors characterized in this work open several perspectives as content recommendation [11], formation of teams of specialists, and the finding of influential users [2].

In the rest of this paper, we first review related work on users activities in UGCs (Section 2). We then present our methodology of data collection, modelling and processing (Section 3). Finally, we present our results on assessing the impact of the social network on marking photos as favorites in Flickr (Section 4), and present the conclusions and future work (Section 5).

2. RELATED WORK

Several studies focus on analyzing users activities in UGCs [3–5, 8]. In general, these studies concentrate on *(i)* analyzing different users' profiles in terms of production and consumption of content [4], *(ii)* modeling the content spread over the social network [5] and *(iii)* analyzing the importance of the network in content dissemination [8].

[1] `www.flickr.com`
[2] `http://blog.flickr.net/en/2011/08/04/6000000000/`

WebMedia'12, October 15–28, 2012, São Paulo/SP, Brazil.

Burke et al. [4] investigated users' activities on Facebook by outlining different users' profiles and how they interact with their network. The study shows that users utilize the system in different ways as regards social activities and the use of the media available. Some users have a more active behavior, i.e., they are more likely to produce than to consume content, while other users exhibit a more passive behavior. In this work we analyze the contacts' network under the user's perspective. We focus on analyzing whether users show a preference for content posted by other users who are closer or distant in the network topology.

Cha et al. [5] analyze how a new content is disseminated within Flickr. The study focuses on the characterization of different patterns in which contents are propagated and on the delineation of a set of characteristics that can be used to classify them. In this study, we focus on analyzing whether users express preferences for contents posted in their network, and, conversely, whether their network is the main consumer of the contents posted by them. We do not analyze content characteristics, we take into account only the nunber of times the content has been marked as favorite.

Doerr et al. [8] analyze users influence in the news aggregator Digg.com. The study shows that some users create networks that have little influence on the content they consume or prefer. Similar behavior has been observed in Twitter [3, 12]. In these UGCs, other behaviors, such as users indirect iterations, are more representative of their influence than the social network that they maintain [3].

In this paper we focus on analyzing the impact of the users' social network on marking photos as favorites in Flickr. This system differs from Digg and Twitter in several ways. In Digg, users aggregate news that have been previously produced and transmitted by other channels, for example, news sites. In Twitter, users often post several small messages without much concern for the quality of the content posted. On the contrary, Flickr users themselves produce the photos posted in the system. Furthermore, users are usually concerned with the quality of the content posted. In many cases the content posted results from a professional or artistic activity.

3. MATERIALS AND METHODS

Our study is based on data collected in Flickr. Regarding users' interface in Flickr, it is important to note that users have access to photos posted by their contacts and those posted by users who are not in their social network. Users' home page exhibits information about photos uploaded or marked as favorites by their contacts. In the same way, users' home exhibits information about photos that are posted and marked as favorites by users who are not their direct contacts, e.g., photos most recently uploaded and photos most marked as favorite in the system.

Data collection.

The data used in this study were collected through the public Flickr's API[3]. The time window considered for data collection was four months. A summary of the data collected is presented in Table 1.

We use *Snowball Sampling* [10] as a crawling method. This method leads to an almost complete collection of data for a particular neighborhood of the social network graph, which is relevant to our analysis. The data collected contain user-user links, and user-item links. User-user links indicate the establishment of contact between a user and another user. User-item links indicate that a user has posted a photo or marked a photo as favorite.

[3]http://www.flickr.com/services/api/

Table 1: Overview of the Flickr Dataset

Period	from Nov 11, 2011 to Feb 29, 2012
#Users	2.241.714
#Photos posted	10.304.604
#favorites marks	29.249.205
#user-user links	20.684.373

Addressed questions.

Our analysis focuses on putting into perspective *(i)* how users are affected by their network, and *(ii)* how they affect the network in which they are included. When we investigate how users are affected by their network, we measure the proportion in which a reference user marks as favorite photos posted by users in their social network. When we investigate how users affect their network, we measure the proportion in which users in a social network mark as favorite the photos posted by a reference user. Thus, taking into account users at different distances in the network topology, we aim at identifying how favorites made and favorites received by each user are distributed along users' network. A scatter plot of the amount of favorites made and favorites received by each user in our dataset is shown in Figure 1.

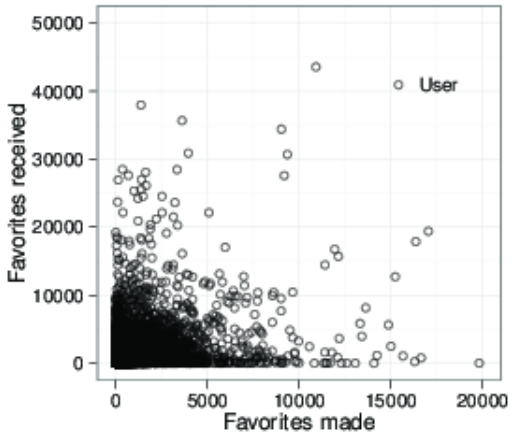

Figure 1: Scatter plot of the number of times that each user has marked as favorite photos posted by other users (favorites made) by the number of times that this user had their photos marked as favorites (favorites received).

Distance between users.

An essential factor in this work is to measure the distance between two users in Flickr. Flickr users' social network can be modeled as a directed graph $G = (U, E)$, where U is the set of users and E the set of directed edges that represent contact between users (user-user links). An example of this type of graph is shown in Figure 2. In this graph, an edge $(u_1, u_2) \in E$ indicates that the user u_1 has the user u_2 as a

contact, and that there is a distance of 0 user from u_1 to u_2. In other words, the distance between a user u_a and a user u_b is the smallest number of users that are between them in the network.

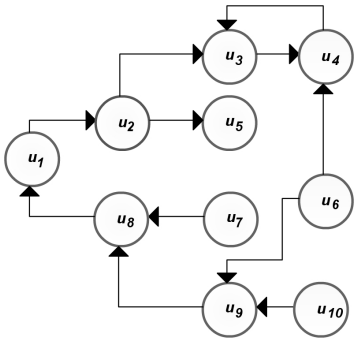

Figure 2: Example of a subgraph of the social network.

In the graph shown in Figure 2, the user u_2 is a contact of the user u_1. The user u_3, in turn, is a contact of the user u_2. In this study, we say that u_3 is at a distance of 1 user from the user u_1. Note that, the user u_7 is not in the network of the user u_1 and he/she cannot be achieved by this user by covering the social network graph. On the other hand, the user u_1 is in the network of the user u_7 and he/she can be achieved by covering the graph. The distance between users can be calculated in a breadth-first search in the graph starting from a reference user up to users on the level n.

We compute users' influence by the favorites their photos receive. For example, when a photo posted by the user u_1 is marked as favorite by the user u_8, we say that the user u_1 has influenced a contact that is at a distance 0 from him/her in the network topology. Otherwise, if the user u_1 marks as favorite a photo posted by the user u_3, we consider that the user u_1 has been influenced by a user who is distant from him/her at 1 user in the social network topology.

Method.

Given the favorites received in photos posted by a user of reference, our method aims at measuring the percentage of favorites made by users from a distance 0 up to a distance n. Similarly, given the favorites made by a reference user, we measure the percentage of such favorites that were made in photos posted by users from distance 0 up to a distance n. To define the reference users, we randomly selected a set of 50 users in our dataset. This number of users proved to be sufficient to obtain results with a satisfactory statistical error for the 95% confidence interval.

Our graph has more than 2 billion vertices (Table 1), and it is necessary to cover it by checking at the same time the photos posted and the photos marked as favorite by each user in different distances starting from the reference users. High computing power is required in this processing. To speed up our experiments, we used the MapReduce programming model to execute our method in parallel using the Hadoop platform. We have made available for download[4] all data used in this work, the scripts used to collect data through the Flickr API, and the MapReduce implementations used to identify the favorites made and favorites received at several distances between users.

4. RESULTS

The results are shown in Figures 3 and 4. In Figure 3 the vertical axis shows the percentages of all photos marked as favorites, and the horizontal axis shows at which distance in the network these photos percentages were uploaded. Each point in the graphic represents the average of the favorites made by 50 reference users. This result shows that 70% of the photos marked as favorite by users were posted by users who are at a distance 0, i.e., those who are their contacts. This indicates that Flickr users consume content shared by users with whom they maintain contact.

Furthermore, the results show that about 20% photos marked as favorite by users were posted by users at a distance of 1 and 2. These photos were disseminated via favorites made by users at the distance 0 like a cascade behavior [5]. Overall, these results show that in Flickr users' network are important in the production of relevant content to their contacts and also in the dissemination of such content.

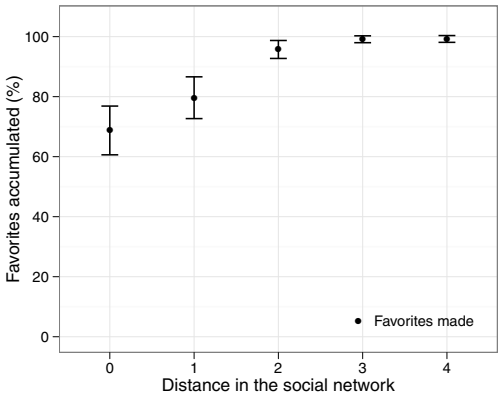

Figure 3: Distance in the social network in which were posted the photos marked as favorites by users.

Figure 4 shows the distances in the social network in which the photos posted by users were marked as favorites. In this figure the vertical axis show the percentage of favorites received in photos posted by reference users, and the horizontal axis shows at the distance in the network other users marked these photos as favorite. Each point in the graphic represents the average of the favorites received in photos posted by 50 users. The result shows that a large part (approximately 60%) of the favorites marks received in photos come from users in the network where they are included.

Implications and Limitations.

Our results indicate that in UGCs like Flickr, users show preference for content posted by their contacts and by contacts of the other users with whom they are connected. The

[4]The data are available at `http://redmine.lsd.ufcg.edu.br/projects/itemsrecommendation/wiki`. More details about implementations are described in the tech report [1].

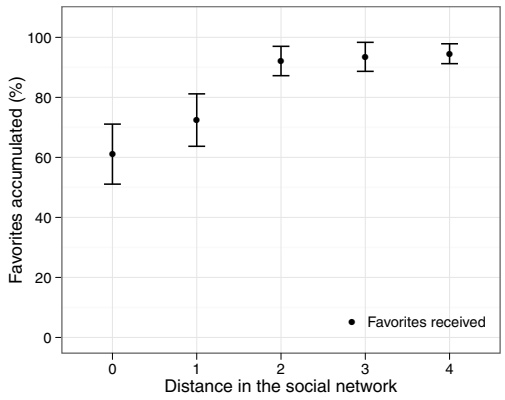

Figure 4: Distance in the social network in which the photos posted by users were marked as favorites.

greater proportion of favorites made by users is in photos posted by users who are their contacts. This result has several implications. For example, regarding content recommendation, this observation indicates that may not be a good approach to recommend photos uploaded closer in users' network, because they usually mark them as favorites even without the help of a recommendation system. However, some users may show more interest by the content posted by their contacts than others. Thus, a limitation of this study is that we have not grouped users in different profiles considering the degree of common interest with their network.

5. CONCLUSION

In this work, we analyze the impact of the users' social network on marking photos as favorites in Flickr. We analyze different distance levels between users in order to identify relationships between the users' preference for the content posted and the users' proximity in the social network topology. Our results indicate that users show preference for content posted by their contacts and by contacts of the other users with whom they are connected. Together, they represent an average of 80% of the favorites made. Moreover, more than 60% of the favorites received in users' photo comes from the network in which they are included.

As future work we intend to group users into different types of social network profiles, considering, for example, the number of contacts they have, the frequency in which they use the system, and the proportion between the amount of photos posted and the amount of favorites received. With this analysis, we intend to identify whether the observed results are more or less significante depending on the user group. Furthermore, given that our results show that users exhibit preference for content posted within their network, we intend to explore this information in a photo recommendation algorithm that find relevant content outside users' network.

Acknowledgments.

Authors are indebted to Nazareno Andrade and Leandro Marinho for their comments and suggestions on earlier versions of this paper. Authors would also like to thank the anonymous referees for their insightful comments. This paper was achieved in cooperation with Hewlett-Packard Brasil Ltda. in the context of the UFCG/HP Analytics Training Program, which uses incentives of the Brazilian Informatics Law (Law no. 8.2.48 of 1991).

6. REFERENCES

[1] L. Alves, M. Maciel, L. Ponciano, and A. Brito. Análise do impacto da rede social na marcação de fotos como favoritas no Flickr. Technical Report UFCG-LSD-2012-01, Distributed System Lab, Universidade Federal de Campina Grande, 2012. Available at `http://www.lsd.ufcg.edu.br/relatorios_tecnicos/TR-2012-01.pdf`.

[2] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi. Online team formation in social networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 839–848. ACM, 2012.

[3] C. Bigonha, T. N. C. Cardoso, M. M. Moro, V. A. F. Almeida, and M. A. Gonçalves. Detecting Evangelists and Detractors on Twitter. In *18th Brazilian Symposium on Multimedia and the Web*, pages 107–114, 2010.

[4] M. Burke, R. Kraut, and C. Marlow. Social capital on facebook: differentiating uses and users. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 571–580. ACM, 2011.

[5] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks*, pages 13–18. ACM, 2008.

[6] L. Chiarandini. Exploration and discovery of user-generated content in large information spaces. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 755–756. ACM, 2012.

[7] E. David and K. Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

[8] C. Doerr, N. Blenn, S. Tang, and P. Van Mieghem. Are friends overrated? a study for the social news aggregator digg.com. *Comput. Commun.*, 35(7):796–809, Apr. 2012.

[9] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(35):75 – 174, 2010.

[10] L. A. Goodman. Snowball sampling. *The Annals of Mathematical Statistics*, 32(1):148–170, 1961.

[11] A. Krohn-Grimberghe, L. Drumond, C. Freudenthaler, and L. Schmidt-Thieme. Multi-relational matrix factorization using bayesian personalized ranking for social network data. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 173–182. ACM, 2012.

[12] A. Silva, H. Valiati, S. Guimarães, and W. M. Jr. From Individual Behavior to Influence Networks: A Case Study on Twitter. In *19th Brazilian Symposium on Multimedia and the Web*, page 8, 2011.

[13] R. van Zwol. Flickr: Who is looking? In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 184–190. IEEE Computer Society, 2007.