Friendship, collaboration and semantics in Flickr: from social interaction to semantic similarity

Andrea Capocci
University "La Sapienza" of Rome, Italy
P.le Aldo Moro, 5, Rome, Italy

Andrea Baldassarri
University "La Sapienza" of Rome, Italy
P.le Aldo Moro, 5, Rome, Italy

Vito D.P. Servedio
University "La Sapienza" of Rome, Italy
P.le Aldo Moro, 5, Rome, Italy

Vittorio Loreto
University "La Sapienza" of Rome, Italy
P.le Aldo Moro, 5, Rome, Italy
ISI Foundation
Villa Gualino, Turin, Italy

ABSTRACT

We study the semantic assortativity in the social networks hosted by the Flickr folksonomy, based both on the contact data and on the group membership data provided by the users. The social network built this way are complex one. Besides, one observes a clear assortativity pattern, stronger than in a suitable null model adopted for a comparison. Nevertheless, such semantical similarity does not appear to develop during the community evolution, but is rather the result of a pre-existing shared background between users.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Systems and Software; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; G.2.2 [**Mathematics of Computing**]: Graph Theory

General Terms

Measurement, Theory

Keywords

folksonomies, semiotics, semiotic dynamics, small worlds

1. INTRODUCTION

The study of online collaborative platforms has recently involved an interdisciplinary scientific community including computer scientists, physicists, linguists and sociologists among others. In particular, collaborative annotation systems have attracted much attention due to their simple structure and the large amount of public data made available through the web. Popular websites such as Delicious, Flickr and CiteULike share the same basic framework: users archive resources online and annotate them by free "tags". Resources may be any piece of digital knowledge, such as web references, photos, scientific references respectively in the mentioned examples. The result of the collective classification by tags is called a "folksonomy".

The structure of these communities are often interpreted as a social network as relationships between pairs of users can be inferred on the basis of their tagging behavior. Typically, pairs of users are connected to build an implicit network if they share a characteristic of interest. For example, in the co-tagging network [1] two users annotating a same set of w resources are assigned a link with weight w.

The Flickr social network, however, exhibits an explicit social structure too. Flickr users declare which peers they like and such peers are called "contacts". Moreover, users with common interests also join "groups" of users to share photos and comments.

Many studies reported in the network theory literature have shown that most self-organized networks, ranging from the genome to the Internet, display peculiar properties such as the short diameter and the fat-tailed distribution of the degree [2]. Besides, social networks are characterized by assortative mixing, i.e. nodes tend to connect to similar peers [11].

Incentives to use similar tags come from multiple sources. For instance, Flickr.com allows to search photos by tags, so that sharing tagging conventions facilitates the navigation of photos. Moreover, tags are also used with strategic purposes [15]. Thus, incentives to use similar sets of tags are strengthened by social interaction. Users in contacts or in the same groups are more exposed to photos and tags coming from their interacting peers, and socially interacting users are prone to share interests or experiences these photos refer to.

Non-trivial correlations in the vocabulary of users in social networks has been already detected [8, 4]. However, it has not been clearly determined whether semantical correlations arise because of the social dynamics taking place within social networks or the shared background knowledge.

2. PREVIOUS WORK

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MSM '10, June 13, 2010, Toronto, Canada.
Copyright 2010 ACM 978-1-4503-0229-6/10/06 ...$10.00.

In the literature, several studies about folksonomies, social networks and their semantic patterns, as separate phenomena, exist. Unlike Flickr, the most commonly studied folksonomies such as Delicious or CiteULike have not embedded any explicit social network among their users until recent times. Thus, little data have been available to analyze how semantics and social connections interact. Instead, Flickr has given users the possibility to join thematic groups and display their personal contacts since its beginnings.

The complex nature of the Flickr directed social networks has been shown in [5, 10, 9] and confirmed by Leskovec *et al.* [7], who have focused on the microscopical mechanisms leading to social link creation. The influence of the Flickr social network on the browsing activity has been explored in [6, 16].

Concerning social networks based on group membership, Pissard and Prieur [12] have studied the groups' link structure and the thematic network structure, based on the usage of same tags by different Flickr users. Their analysis shows that the density of a group does not imply a strong common thematic interaction.

The semantic similarity between contact users' tag clouds has been first explored by Marlow *et al.* [8], showing that the distribution of the Jaccard index (i.e., the ratio between the intersection and the union of the tag sets employed by two users) has a larger mean value and variance when measured over personal contacts than on random pairs of users. A more rigorous approach has been followed in [14], where the semantic similarity (tag cloud cosine similarity) has been measured as a function of the distance between users in the contact social network, showing that topological and semantical closeness are positively correlated.

3. QUANTITIES OF INTEREST

The notion of semantic similarity arises in many different contexts, although the methods to measure it vary widely. For instance, in classical linguistics literature, semantic similarity is measured between concepts or words in order to build hierarchical taxonomies [13]. Here, we measure the similarity between the sets of tag assignments of individual users, described by *tag clouds*. A tag cloud is mathematically represented by a vector $t^u = (t_1^u, ..., t_N^u)$ where t_i^u is the occurrence of tag i in the set of tags employed by user u.

To take into account the different weights of tags in a tag cloud, one computes the cosine similarity - the cosine of the angle between the vectors representing two tag clouds [3, 14]. The cosine similarity $C(u, w)$ between users u's and w's tag clouds is defined as

$$C(u,w) = \frac{t^u \cdot t^w}{|t^u||t^w|}, \qquad (1)$$

where the norm reads $|t| = \sqrt{t \cdot t}$. Cosine similarity takes its maximum value 1 if the tag clouds are identical, and takes its minimum value (null) if two tag clouds have no common tag.

4. DATASETS

The dataset we analyze here covers one year (2006) of Flickr activity, that is, a list of 109294825 tag assignments. For each tag assignment, the timestamp is available, along with the author, the resource and the tags. Beside the posting activity, we also have crawled the explicit Flickr social networks: for each users, his contacts and group memberships at the end of 2006 are available. The contact data allow us to build a directed network, where edges go from users to their contacts.

5. CONTACT-BASED SOCIAL NETWORK

Flickr contacts allow to build two types of social network. A directed version takes into account all contacts, assuming a directed link from a users and his/her friends. A second, undirected version, takes into account only the mutual contacts.

As it has already been computed in the literature, the directed contact network is a scale–free one in both the in-degree and the out-degree with a very similar decay exponent, as shown in figure 1. A similar conclusion can be drawn for the undirected mutual contact network.

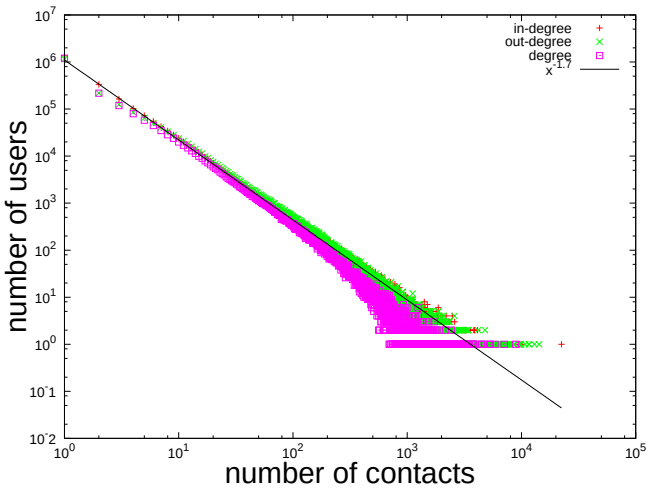

Figure 1: The in-degree and the out-degree distribution measured in the Flickr directed contact network, and the degree distribution measured in the Flickr undirected mutual contact network.

6. GROUP-BASED SOCIAL NETWORK

We use the group structure, too, as a method for identifying social networks. In this case, a connection between two users means that they are members of a same group. Hence, such group-based social networks are undirected ones. Flickr groups form an extremely heterogeneous set. As shown in figure 2, the size of Flickr groups displays a broad distribution, with groups composed by hundreds of thousands users. Therefore, one can hardly define something such as a "typical" Flickr group.

Group size does not only affects the connectivity in the corresponding social network, but also the link strength. Pissard *et al.* [12] have suggested that members of large groups are more loosely connected than members of small groups. To explore such property, we have derived several social networks from the original group data by neglecting memberships to groups with more than S_{max} members, for different values of S_{max}. As reported in figure 3, the average tag cloud cosine similarity displays an approximately exponential decay as S_{max} increases, confirming the findings of Pissard *et al.* [12], as expected.

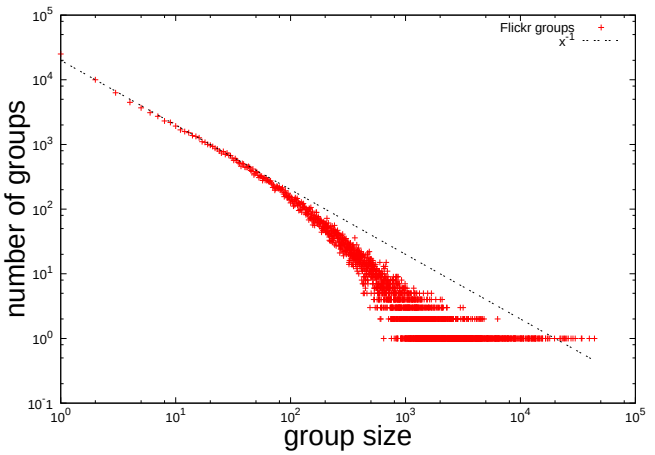

Figure 2: The distribution of the size of Flickr groups.

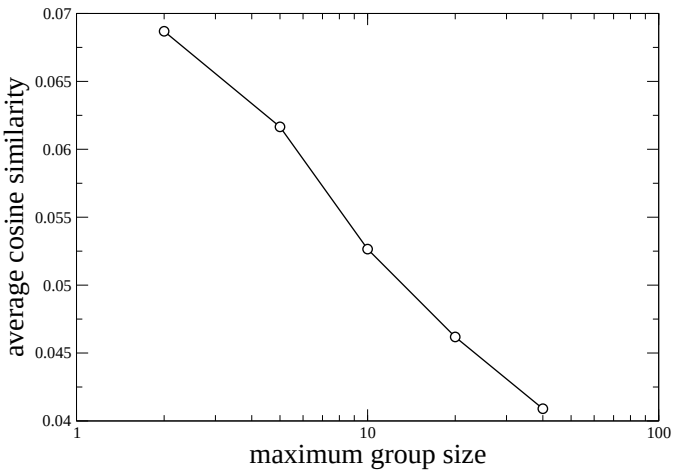

Figure 3: The average tag cloud cosine similarity computed in groups with less than S_{max} members.

Therefore, in the following group averages will be characterized by the group size, since similarity assessments vary strongly as a function of it. By $G-n$, we will refer to the social network composed by users sharing at least one common membership to groups of maximum size n.

We have measured the average nearest–neighbors semantic similarity for different social networks and possible definitions of it. Since we are dealing with a strongly off-equilibrium systems, drawing conclusions for time-dependent quantities is risky. The growth mentioned above may be the result of the skewed distribution of tag frequencies. Hence, one has to compare the observed signal with the same measurement performed in a suitable null model. The null model we adopt is based on the randomization of the tag stream, and leaves the social network unchanged, as described in [14]. In each time interval, we build the global list of tags with their multiplicities, where each tag appears the total number of times it has been used in the time interval. Then, for each user, each distinct tag is replaced by a random one drawn with uniform probability from the global list of tags, which is assigned the frequency of the real tag.

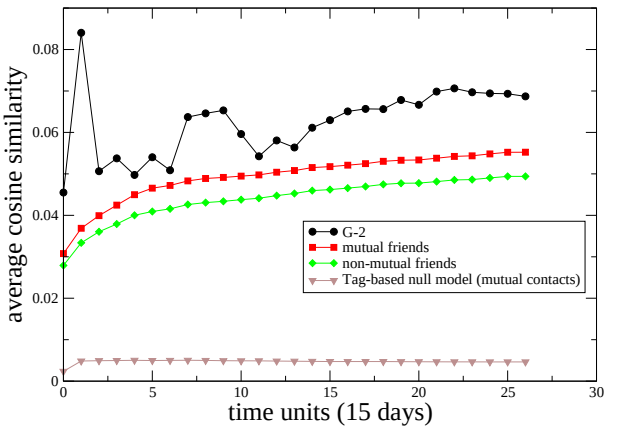

Figure 4: The time evolution of the cumulative cosine similarity for different social networks and the null model.

7. ALIGNMENT DYNAMICS

We focus now our attention on the process through which such similarity is developed. The observed semantic alignment could take place because of two possible mechanisms at work. In a first hypothesis, the similarity between users develops because of their social interaction, as in an imitative process. In the second one, the social interaction is the signature of a pre-existing shared background: in the latter case, the semantic similarity is not a consequence, but rather a premise of social connection.

In figure 4, we show how the semantic similarity evolves in time. We have computed the average tag cloud similarity over all pairs of neighbor users in the different social networks as a function of time. The semantic similarity is computed by taking into account the tag cloud of each user from the initial time ("cumulative" tag cloud) but the average is computed only on users who have been active in a time window of a given length.

In all examined networks, the average neighbors' tag cloud similarity increases in time, whereas in the considered null models the same quantity remains well below. The growth of the similarity changes its rate and stabilizes roughly after a period of about three months, where it reaches approximately its stationary value.

To understand whether such alignment arises because of imitative tagging, we have measured the "snapshot" average tag cloud similarity (whose time evolution is reported in figure 5) – the similarity referring only to the tagging activity that took place within the last time interval. The "snapshot" semantic similarity does not increase in time, but remains rather stable. So, the growing similarity observed in figure 4 is not due to a synchronization of the tag usage in neighbors, but rather to an asynchronous usage of a similar set of tags.

In both the "cumulative" and "snapshot" cases, however, the tag cloud alignment seems to be a genuine signature of a semantical phenomena taking place, since the alignment measured in the corresponding null model remains well below the observed values.

As a result of the above observation, social interaction and semantic similarity appear indeed to be positively correlated in the Flickr social netowrk, but this correlation is determined more by the shared background knowledge and

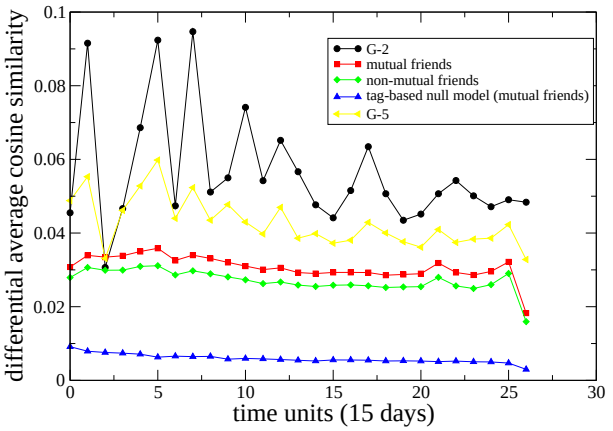

Figure 5: The time evolution of the "snapshot" average nearest neighbors' cosine similarity for different social networks and the null model.

does not emerges during time because of the social dynamics taking place within the network.

8. CONCLUSIONS

We have analyzed a dataset reporting the individual and social activity of users in the Flickr network during a period of one year (2006) to investigate how social relations in virtual communities influence semantic similarity between users. We have shown that semantic similarity is larger between socially interacting users and reaches its maximum between group starters, i.e. in groups with only two users.

To uncover the semantic properties of the Flickr social network we have compared it to a suitable null model, based on a random re-assignment of tags. The null models display a pattern of lower similarity with respect to the real social networks, as expected. The dynamics of the tag cloud alignment, however, shows that the observed similarity is not the result of the social interaction within the Flickr groups; rather, it is determined by the existence of a shared background knowledge, and the interaction taking place in the Flickr social network appears to have little effect on the semantics of the folksonomy.

9. ACKNOWLEDGMENTS

This research was supported by the TAGora project (FP6-IST5-34721) funded by the Future and Emerging Technologies program (IST-FET) of the European Commission. We acknowledge useful discussions with Alain Barrat and Ciro Cattuto.

10. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, volume 22, pages 207–216, New York, NY, USA, June 1993. ACM Press.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97, Jan 2002.

[3] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic analysis of tag similarity measures in collaborative tagging systems, May 2008.

[4] W.-T. Fu, T. G. Kannampallil, and R. Kang. A semantic imitation model of social tag choices. In *CSE '09: Proceedings of the 2009 International Conference on Computational Science and Engineering*, pages 66–73, Washington, DC, USA, 2009. IEEE Computer Society.

[5] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM Press.

[6] K. Lerman and L. Jones. Social browsing on flickr. In *Proceedings of International Conference on Weblogs and Social Media (ICWSM)*, March 2007.

[7] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, New York, NY, USA, 2008. ACM.

[8] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*, pages 31–40, 2006.

[9] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the flickr social network. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 25–30, New York, NY, USA, 2008. ACM.

[10] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *In Proceedings of the 5th ACM/USENIX Internet Measurement Conference (IMC'07*, 2007.

[11] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, Oct 2002.

[12] N. Pissard and C. Prieur. Thematic vs. social networks in web 2.0 communities: A case study on flickr groups. In *Proc. of Algotel Conference, 2007. http://hal.inria.fr/inria-00176954/en*, 2007.

[13] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[14] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. Folks in folksonomies: social link prediction from shared metadata. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 271–280, New York, NY, USA, 2010. ACM.

[15] L. Steels and E. Tisselli. Social tagging in community memories. In *Proceedings of AAAI Symposium on Social Information Processing*, 2008.

[16] R. van Zwol. Flickr: Who is looking? In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 184–190, Washington, DC, USA, 2007. IEEE Computer Society.