HAL
open science

Friendship prediction and homophily in social media

Luca Maria Aiello, Alain Barrat, Rossano Schifanella, C. Cattuto, Benjamin Markines, Filippo Menczer

To cite this version:

Luca Maria Aiello, Alain Barrat, Rossano Schifanella, C. Cattuto, Benjamin Markines, et al.. Friendship prediction and homophily in social media. ACM Transactions on the Web, 2012, 6 (2), pp.9. 10.1145/2180861.2180866 . hal-00718085

HAL Id: hal-00718085

https://hal.science/hal-00718085v1

Submitted on 16 Jul 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Friendship prediction and homophily in social media

LUCA MARIA AIELLO
Department of Computer Science, University of Turin, Italy
ALAIN BARRAT
Centre de Physique Théorique (CNRS UMR 6207), Marseille, France
Complex Networks and Systems Laboratory, ISI Foundation, Turin, Italy
ROSSANO SCHIFANELLA
Department of Computer Science, University of Turin, Italy
CIRO CATTUTO
Complex Networks and Systems Laboratory, ISI Foundation, Turin, Italy
BENJAMIN MARKINES
FILIPPO MENCZER
School of Informatics and Computing, Indiana University, Bloomington, IN, USA

Web 2.0 applications have attracted considerable attention because their open-ended nature allows users to create lightweight semantic scaffolding to organize and share content. To date, the interplay of the social and topical components of social media has been only partially explored. Here we study the presence of homophily in three systems that combine tagging of social media with online social networks. We find a substantial level of topical similarity among users who lie close to each other in the social network. We introduce a null model that preserves user activity while removing local correlations, allowing us to disentangle the actual local similarity between users from statistical effects due to the assortative mixing of user activity and centrality in the social network. This analysis suggests that users with similar interests are more likely to be friends, and therefore topical similarity measures among users based solely on their annotation metadata should be predictive of social links. We test this hypothesis on several datasets, confirming that social networks constructed from topical similarity capture actual friendship accurately.

Categories and Subject Descriptors: H.1.2 [**Information Systems**]: Models and Principles—*Human information processing*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Collaborative computing, Web-based interaction*

General Terms: Algorithms, Experimentation, Measurement

Additional Key Words and Phrases: Web 2.0, social media, folksonomies, collaborative tagging, social network, link prediction, topical similarity, Maximum Information Path

---

Contact author: Luca Maria Aiello, aiello@di.unito.it.
The present paper is an extended version of a previously published conference paper [Schifanella et al. 2010].
Permission to make digital/hard copy of all or part of this material without fee for personal or classroom use provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.
© 20 ACM 0000-0000/20/0000-0001 $5.00

1. INTRODUCTION

Online social networking systems, together with online systems for content organization and sharing, entangle cognitive, behavioral, and social aspects of a user community through an underlying technological platform. The resulting "ecosystems" provide new possibilities to mine and investigate the various processes at play in the interactions of individuals, and to study the ways in which users relate with the information they share.

Key open questions deal with understanding the concepts of similarity and influence, tracking the emergence of shared semantics, and determining the interplay between social proximity and shared topical interests among users. The emergence, spreading, and stability of any shared concept depend critically on the above factors. As observed by danah boyd [2009],

> *"In a networked world, people connect to people like themselves. What flows across the network flows through edges of similarity. The ability to connect to others like us allows us to flow information across space and time in impressively new ways, but there's also a downside. [...] In a world of networked media, it's easy to not get access to views from people who think from a different perspective. Information can and does flow in ways that create and reinforce social divides. Democratic philosophy depends on shared informational structures, but the combination of self-segmentation and networked information flow means that we lose the common rhetorical ground through which we can converse."*

We see a pressing need for a data-driven investigation of these issues. Social media supporting tagging are especially interesting in this respect because they stimulate users to provide light-weight semantic annotations in the form of freely chosen terms [Golder and Huberman 2006]. Social annotations based on tags are valuable for research because they externalize the three-way relation between users, items of interest (resources), and metadata (tags). Usage patterns of tags can be employed to monitor interests, track user attention, and investigate the emergence and spread of shared concepts through a user community. Moreover, several "Web 2.0" resource organization systems support explicit representations of social links between users, making an objective definition of social proximity available. They also combine several aspects of user activity, such as exposing resources, tagging items, belonging to discussion groups, and relating to other users.

In this paper, we consider three different online social systems: Flickr, Last.fm, and aNobii. In these systems, users expose resources (pictures, songs/artists, and books, respectively), form social networks **and tag items producing social classification of data commonly called *folksonomies***.

The three systems strongly differ by size, category of exposed resources, and the precise ways in which users tag resources and relate to each other. Taking advantage of the datasets built from the three systems, we address the following issue: *How does the similarity between user profiles relate to their proximity on the social network?* More precisely, are neighboring users more similar, both in the amount of activity they devote to the system, and in the content of their activity, than users who lie further apart in the social network? If so, how does this local

similarity fade when the distance on the social graph increases? And can we predict the existence of social links from knowledge of the similarity among user profiles?

In the remainder of this paper, after a brief description of related work in Section 2 and of our datasets in Section 3, we provide in Section 4 a thorough analysis along several axes. This data analysis highlights the heterogeneity of user activities and the correlations in the various metrics measuring the different activities of a single user. We also show the existence of non-trivial mixing patterns: the amount of different activities of a user is correlated with her neighbours'. Section 4 exposes the substantial level of several types of topical similarity that exist among users who are close to each other in the social network. In Section 5, we evaluate the performance of predictors of online social links based on the similarity of user profiles. We consider a number of topical similarity measures from the literature. Scalable similarity measures, such as Maximum Information Path, proposed by some of the authors, are among those achieving the best predictive performance. The role of language communities in these predictions is investigated in Section 6.

2. RELATED WORK

In prior work we explored the correlation between proximity in an online social network and topical similarity and we analyzed the extent to which similarity among users based only on tagging information can be an accurate predictor of social ties [Schifanella et al. 2010]. We analyzed samples from the Flickr and Last.fm social networks. The present paper expands on our prior work both in breadth and in depth. First, we extend our analysis to the aNobii network, which is very different in both its size and the types of items exposed. Second, we widen the social features analysis by including the predictive potential of groups and user libraries (i.e., collections of items like artists or books). Third, we compare the predictive power of topical similarity measures with more sophisticated baselines from the industry. Lastly, we investigate the influence that confounding aspects like user language can have on the link prediction performance.

Similarity between the members of social groups, or between individuals sharing a social link, is known as homophily in the social networks literature, and has long been observed and studied [McPherson et al. 2001]. Homophily phenomena can be present because of selection mechanisms (individuals create social links preferentially with other individuals sharing a certain degree of similarity), but also because of social influence (linked individuals influence each other and become more similar), two effects that are often confounded and actually difficult to disentangle [Leenders 1997; Aral et al. 2009; Shalizi and Thomas 2010]. Interestingly, the everincreasing availability of data sets concerning online social networks have made such networks ideal laboratories for testing and quantifying such social phenomena and theories (see e.g. [Crandall et al. 2008; Aiello et al. 2010; Szell et al. 2010; ?]).

We find indeed in the literature several studies on the evolution of online social systems and correlations between different user features. Leskovec and Horwitz [2008] present a study on the Microsoft Messenger network, showing correlation between user profile information and communication patterns. Evolutionary pat-

terns in the Flickr social network have been studied by Kumar et al. [2006] and Mislove et al. [2007; 2008]. Marlow et al. [2006] perform a quantitative study on the tag usage in Flickr. They discuss the heterogeneity of tagging patterns and perform a preliminary analysis of vocabulary overlap between pairs of users. Their analysis shows that neighbors in the social graph have a higher vocabulary overlap, on average. However, no assessment is made of biases that could be responsible for the reported observation: here we explore such biases. The role of groups as coordination tools in Flickr is investigated by Prieur et al. [2008]. They also point out a strict relation between the density of the social network and the density of the network of tag co-usage among the group members. Leskovec et al. [2008] perform a comparative study on the microscopic evolutionary dynamics between several social networks, in which a special emphasis is placed on the arrival process of new nodes and on the dynamics of attachment. Influence of social contacts on browsing patterns in Flickr has been analyzed by Lerman and Jones [2007] and van Zwol [2007], who provide insights into the activity patterns of users. Correlation between topical overlap among user interests in tagging systems and other indicators of social behavior is explored by Santos-Neto et al. [2009]. They consider CiteULike and Connotea systems, which both lack an explicit social network component, so they look at collaboration relations determined by the participation in the same discussion group.

Predicting the presence of a link between two entities in a network is one of the major challenges in link mining [Getoor and Diehl 2005]. A common approach to the link prediction problem is to infer ties from the structural properties of the social network. Liben-Nowell and Kleinberg [2003] discuss several notions of node similarity based on social graph structural features for link prediction. Prediction of future links in a question-answering bulletin board service is performed by Murata and Moriyasu [2007]. Here, network proximity scores calculated from local topological information are assigned to pairs of nodes, as the proximity values are shown to be accurate predictors of future links. Huan [2006] defines a cycle formation model for social graphs that relates the probability of the presence of a link with its ability to form cycles. The parameters of the model are estimated using the generalized clustering coefficients of the network. The power of the model is evaluated on the Enron email dataset. Another probabilistic network evolution model aimed at link prediction is proposed by Kashima and Abe [2006]. The idea is that links appear in the network due to a copying process where status labels associated to edges are copied from one node to another with a probability that is dependent on the relative topological position of the two nodes. **Clauset et al. [2008] present a hierarchical decomposition algorithm for network clustering which can also be applied to predict missing interactions in networks. The generated graph-dendrograms determine the probability of connection for every pair of vertices. Links are predicted between pairs that have high probability of connection within the hierarchical random graphs but that are unconnected in the observed network. This technique is tested with good results on several small-size networks.**

Another line of works focuses on link detection through supervised learning methods trained on the topological features of graph [Popescul

et al. 2003; Hasan et al. 2006]. Prediction of the sign of an existing link in friends-foes social networks (e.g. Slashdot Zoo) using a machine learning approach is presented by Leskovec et al. [?]. They use a logistic regression classifier trained with two classes of topological features: node degree and triadic closure.

Link prediction can also be based on features that describe user profiles, based on the principle that people with similar tastes are more likely to establish social contacts. Caragea et al. [2009] study the interplay between social network structure and user profile features in the prediction of social ties. The paper proposes an ontology-based classification of user features and shows that the semantics captured by the ontology can effectively improve the performance of a topology-based machine learning classifier for link prediction. Li et al. [2008] propose a system to cluster users with similar topical interests. Starting from a Delicious dataset, the system extracts implicit relations between groups of users based on the similarity of their tag vocabulary. Although the authors do not refine the interest clusters in a set of binary social connections, the approach is related to the feature-based link prediction task. **Leroy et. al. [Leroy et al. 2010] leverage the group membership information from Flickr to build a probabilistic graph useful to detect the hidden social graph.** Mislove et al. [2010] explore the complementary question: can we predict topical similarity from the social network? Again, here we discuss the role of global correlation in biasing such prediction.

Even if the majority of papers is focused on link prediction on simple (directed or undirected) graphs, techniques have been developed also for different kind of networks like weighted networks [Lü and Zhou 2009], bipartite networks [Dunlavy et al. 2010; Benchettara et al. 2010; Kunegis et al. 2010] and signed social graphs [?]. Finally, some approaches based on probabilistic models such as relational Markov networks [Taskar et al. 2003] and probabilistic relational models [Getoor et al. 2003] deserve to be cited. However, these approaches have not been extensively tested on real-world datasets.

A comprehensive survey on link prediction techniques has been recently drawn by Lü and Zhou [2010]; authors compare several structural similarity metrics for link prediction in terms of accuracy and computational efficiency.

In our previous work [Markines et al. 2008; Markines et al. 2009; Markines and Menczer 2009] we made a systematic analysis of a broad range of semantic similarity measures that can be applied to the three-dimensional folksonomy space to extract similarity networks of tags, resources, or users. Here, we use such measures to perform link prediction based on the folksonomy information.

3. DATASETS

In the following, we report on the main features of our datasets and we describe the data retrieval methods we used to build them. For each dataset, we collected at least the information about the social network, the tag assignments, and the group affiliations. A summary about the size of the quantities involved for each dataset is reported in Table I.

3.1 Flickr

We collected the tagging information about the pictures uploaded in Flickr between January 2004 and January 2006 by means of API methods (`flickr.com/api`). The crawling effort was distributed by splitting the above time interval into smaller time windows to be crawled independently. A global tag knowledge base, initialized with a minimal set, was shared between parallel crawlers. Crawlers issued search queries limited to their specific time interval to retrieve information about photos marked with the tags stored in the common database. New tags were added to the shared database as they were discovered by individual crawlers.

Separate crawls were made to explore group affiliations and the social network. In Flickr jargon, social ties are called *contacts*; they are *directed* and do not require acceptance by the linked user. The overall crawl was performed during the first half of 2007.

Our analysis will focus on the network of about 130 thousand users for whom we have tag, group, and contact information.

3.2 Last.fm

In Last.fm, each user is linked to *friends* through undirected links that are established given the consent of both endpoints. Users also have a public list of *neighbors*, computed by the system as recommendations for potential new friendship contacts. An affinity value, ranging from 0 to 1, is also assigned to each member of the neighbor set. Users can annotate songs, artists or albums with tags, and can create or join groups. Users also have a public *library*, i.e., a list of the artists they have listened to. User profile information includes an optional geographic specification at the country level.

We used both API calls (`last.fm/api`) and web crawling methods to build the dataset. The API can be used to retrieve user profiles, friendships and neighborhood relationships and a list of the 50 top artists in the user library (i.e., those with the highest playcount). The API does not allow for the collection of a user's complete activity and group affiliation information, so we extracted the (*user, item, tag*) triples and the group membership relations via web crawling and scraping. The user set we consider was selected by a BFS crawl of the friendship network. The crawls took place in January 2010. We started from three randomly chosen users and for each of them we performed a crawl up to those nodes that resided 4 hops away. The corresponding snapshots include approximatively 100 thousand users each, with an overlap of about 20% between them. Since we found that the results of our analysis are consistent across the three samples, we report the findings for a single representative one.

Recently, the Last.fm API was extended with a similarity function, called *tasteometer*, which, given in input a pair of users or artists, returns an affinity score ranging from 0 to 1. This value is different from the one provided by the neighborhood score and, most of all, it can be computed for *any* pair of users or artists. Jointly with the crawling activity, we retrieved the tasteometer values for a large set of user pairs to compare the performance of our tag-based similarity functions in the link prediction task with the performance of the tasteometer. Further details are given in Section 5.

Table I. Dataset statistics

Dataset	Users	Triples	Tags	Tagged items	Groups
Flickr	130,840	90,723,412	1,420,656	20,599,583	92,301
Last.fm	90,049	6,971,166	194,763	894,615	69,306
aNobii	86,800	5,378,190	143,182	918,181	3,581

3.3 aNobii

Users in aNobii (`anobii.com`) fill their digital book collections with titles selected from the public aNobii book database, which contains the metadata (such as publication year, authors etc.) of about 20 millions different publications, written in 49 different languages. Each personal book collection is partitioned into a *library*, which is a set of titles that the user is reading or has already read, and a *wishlist* that lists the books that the user wants to read in the near future. Books in the user collection can be annotated with arbitrary tags. Since books in libraries form the vast majority of the overall book collections, here we focus mainly on books from libraries.

Users can also provide public information about their profile, such as gender, age, marital status, and a detailed specification of their geographic location including country and hometown. Affiliation with thematic, user-generated groups is also possible.

Two different types of social ties can be established between users: *friendship* and *neighborhood*. The aNobii website suggests that people should be friends if they know each other in real life. Users should establish neighborhood ties with people who have a library they consider interesting. Surprisingly, although these two types of social links are formally different, they are equivalent from a structural point of view. In fact, both are *directed* and can be created without any consent of the linked user, who is not even notified when a new incoming tie is established. Furthermore, both links activate a monitoring on the linked user's library that triggers notifications on library updates. Given this strong structural similarity, and since the two types of links are *mutually exclusive*, in the following we deal with the *union* between friendship and neighborhood networks and we generically refer to the union network as the aNobii social network.

We crawled the aNobii network in December 2009 starting from a random seed of users and following the social links in a forward BFS fashion. We explored the entire giant strongly connected component and the out component of the social network, for a total of 86,800 users. We collected each user's profile information, group affiliations, library, and tag assignments through web scraping.

4. DATA ANALYSIS

In most folksonomies, the activity of users has many facets. In Flickr, for instance, users can upload pictures and tag them, participate in groups, and comment on photos. In Last.fm, users can listen to music, tag songs according to the songs' characteristics or the user's tastes. In aNobii, users can add books to their libraries, tag them, join groups, and create a list of books they wish to read.

Since social networks are explicitly built by users, we can also consider the number of friendship relations to be a measure of activity in each folksonomy we con-

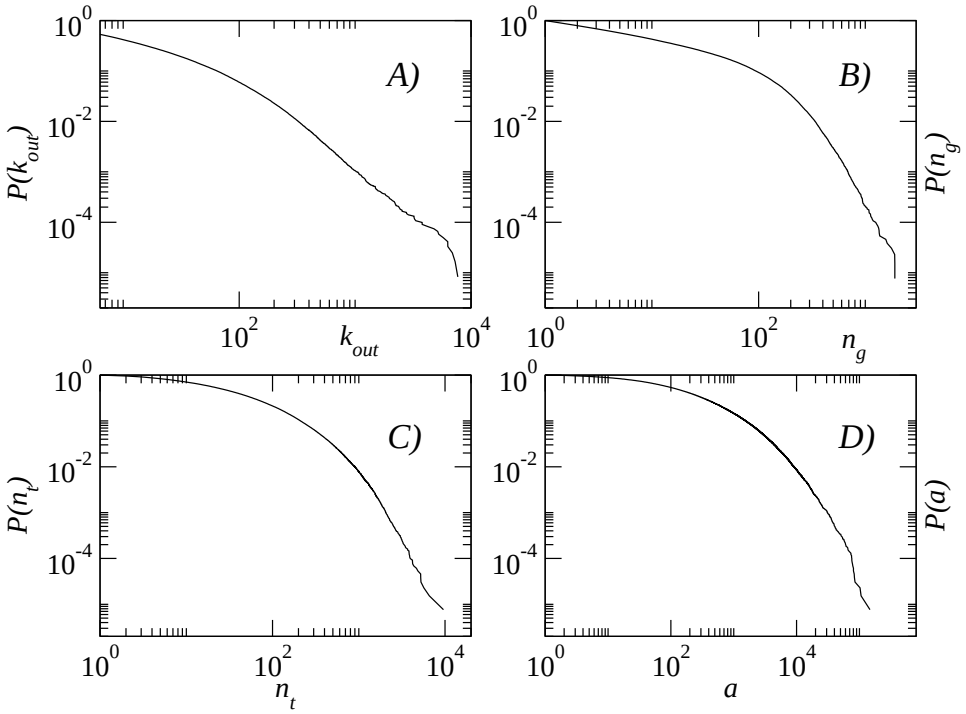

Fig. 1. Flickr complementary cumulative distributions of (A) the number k_{out} of neighbors of a user, (B) the number n_g of groups of which a user is a member, (C) the number n_t of distinct tags per user, and (D) the number a of tag assignments per user.

sider. When links are directed, the out-degree can be considered a more significant measure of activity, while the in-degree measures popularity.

In this section, we first analyze the activity patterns of individual users, and show their considerable heterogeneity. We also investigate the correlations between various activity indicators.

4.1 Heterogeneity and Correlations

Let us first focus on the diversity between users. Figures 1 and 2 show the distributions of the number of neighbors in the social network and the probabilities of finding a user with a given number n_t of distinct tags in her vocabulary, a total tagging activity a, belonging to n_g groups, and having (for aNobii) n_b and n_w books in her library and wishlist, respectively.

All these distributions are broad, spanning multiple orders of magnitude, showing that the activity patterns of users are highly heterogeneous. For each activity measure, most users have little activity, but certain users are on the contrary extremely active, and all intermediate values are represented. No characteristic or "typical" value of the activity can be sensibly defined as evident from a standard deviation that is orders of magnitude larger than the average, for each activity measure.

Given this high level of disparity between users, a natural question arises about the correlations between the different types of activity: do users who have many

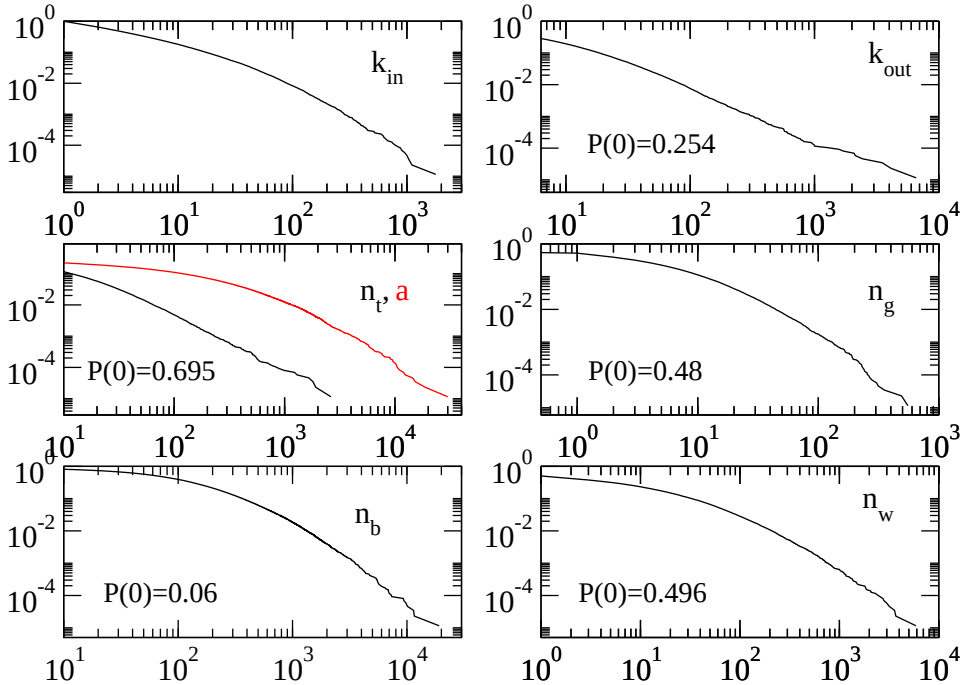

Fig. 2. Complementary cumulative distributions of the measures of activity of aNobii users: in-degree k_{in} and out-degree k_{out} in the social network, number of distinct tags n_t and total tagging activity a (total number of tags in a user's page), number of group memberships n_g, number of books in a user library n_b and in a user wishlist n_w.

neighbors also use many tags, belong to many groups, and so on? The simplest way to examine this issue is to compute the average activity of a type for users having a certain value of another activity type. For instance, we can measure the average number of distinct tags for users having k neighbors in the social network:

$$\langle n_t(k) \rangle = \frac{1}{|u : k_u = k|} \sum_{u:k_u=k} n_t^u \,, \tag{1}$$

where n_t^u is the number of distinct tags of user u. As shown in Figure 3, all types of activity have a clearly increasing trend for increasing values of the out-degree; users who have more contacts in the social network tend also to be more active in terms of tags and groups. Overall, the various activity metrics are all positively correlated with one another. For instance, in Flickr, the Pearson correlation coefficients are: 0.349 between k and n_t, 0.482 between k and n_g, 0.268 between k and a, 0.429 between n_t and n_g, 0.753 between n_t and a, and 0.304 between n_g and a.

Despite these correlations, large fluctuations are still present. First, the strong fluctuations at large degree values are due to the smaller number of highly-connected nodes over which the average is performed. Notably, users with a large number of social contacts but using very few tags and belonging to very few groups can be observed. We can investigate in more detail the degree of correlation between activity types through the conditional probabilities of the type $P(n_t|k)$, i.e., the

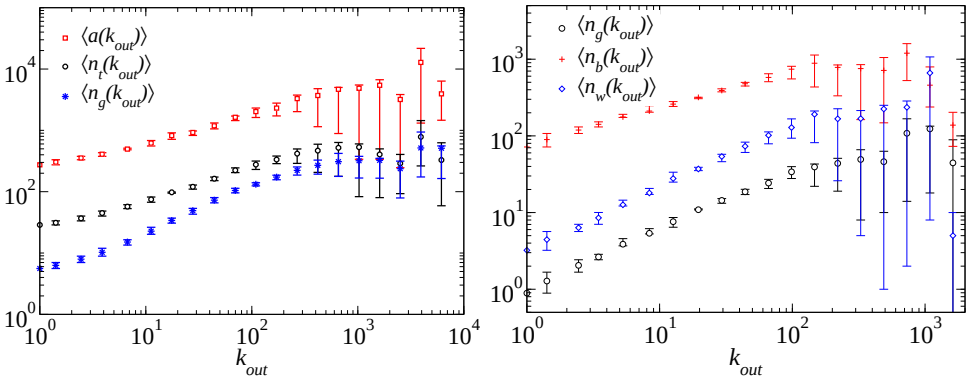

Fig. 3. Left: Average number of distinct tags ($\langle n_t \rangle$), of groups ($\langle n_g \rangle$), and of tag assignments ($\langle a \rangle$) of users having k_{out} out-neighbors in the Flickr social network. Right: Correlations between the activity of aNobii users and their number of declared friends and neighbors: group memberships $\langle n_g \rangle$, library $\langle n_b \rangle$ and wishlist sizes $\langle n_w \rangle$, averaged over users with k_{out} out-links, vs k_{out}. The data has been log-binned: the symbols indicate the average, and the errorbars the 25 and 75 percentiles for each bin.

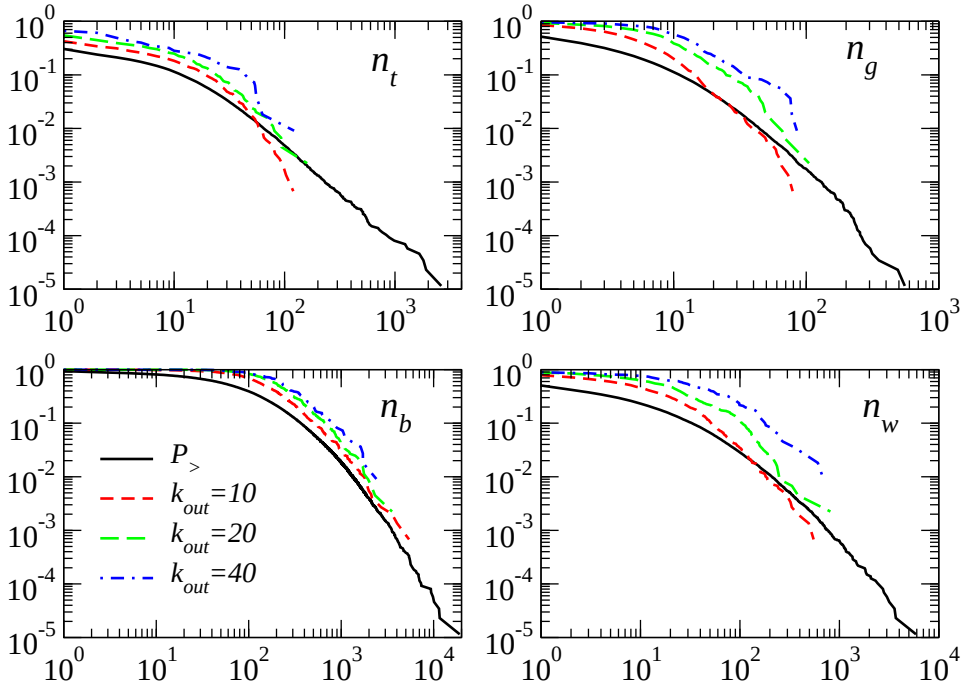

Fig. 4. Complementary cumulative conditional distributions $P(n|k_{out})$ in aNobii, where n is the number of tags n_t, of groups n_g, of books n_b, and wishlist size n_w, compared with the global cumulative distributions $P_>$ (black lines). Even among the subset of users with a given k_{out}, a strong disparity is still observed in the amount of activity.

probability for a user to have n_t tags, knowing that she has k neighbors in the social network, where the average $\langle n_t(k) \rangle$ is simply the first moment of this conditional distribution. As shown for some examples in Figure 4, these distributions, although narrower than the distributions shown in Figures 1 and 2, remain broad. This shows that, despite the strong correlations observed, users with a given activity level in the social network remain quite heterogeneous.

4.2 Mixing patterns

While the previous analysis concerns the correlations between the diverse activity levels of a single user, another important question concerns the correlations between the activity metrics of users who are linked in the social network. This is a long-standing problem in social sciences, ecology and epidemiology: a typical pattern, referred to as "assortative mixing," describes the tendency of nodes of a network (here, the users), to be linked to other nodes with similar properties [Newman 2003]. This tendency appears intuitive in the context of a social network [Newman 2002; Newman and Park 2003], where one expects individuals to be preferentially connected with other individuals sharing the same interests, and the property is then also called "homophily" [McPherson et al. 2001]. Likewise, it is possible to define a "disassortative mixing" pattern whenever the elements of the network tend to link to nodes that have different properties. Mixing patterns can in fact be defined with respect to any property of the nodes. In the present case, we can characterize the mixing patterns concerning various activity types.

In the case of large scale networks, the most commonly investigated mixing pattern involves the degree (number of neighbors) of nodes. This type of mixing concerns the likelihood that users with a given number of neighbors connect with users with similar degree. This property is investigated by computing multi-point degree correlation functions. The correlation between the degrees of connected users are measured by the conditional probability $P(k'|k)$ that a given user with degree k is connected to a user of degree k'. Such a quantity is highly affected by statistical fluctuations, so a more commonly used measure is given by the average nearest neighbors degree of a user u,

$$k_{nn}^u = \frac{1}{k_u} \sum_{v \in \mathcal{V}(u)} k_v \ , \tag{2}$$

where the sum runs over the set $\mathcal{V}(u)$ of neighbors of u. To characterize mixing patterns with respect to nodes' degrees, a convenient measure can be built on top of k_{nn}^u by averaging over all nodes u that have a given degree k [Pastor-Satorras et al. 2001; Vázquez et al. 2002]:

$$k_{nn}(k) = \frac{1}{|u : k_u = k|} \sum_{u:k_u=k} k_{nn}^u \ , \tag{3}$$

which turns out to be the first moment of $P(k'|k)$.

In the case of folksonomies, since each user is endowed with several properties characterizing his activity, it is interesting to characterize mixing patterns with respect to each of these properties. To this end, we generalize the average nearest neighbors degree presented above, and define for each user u the average number

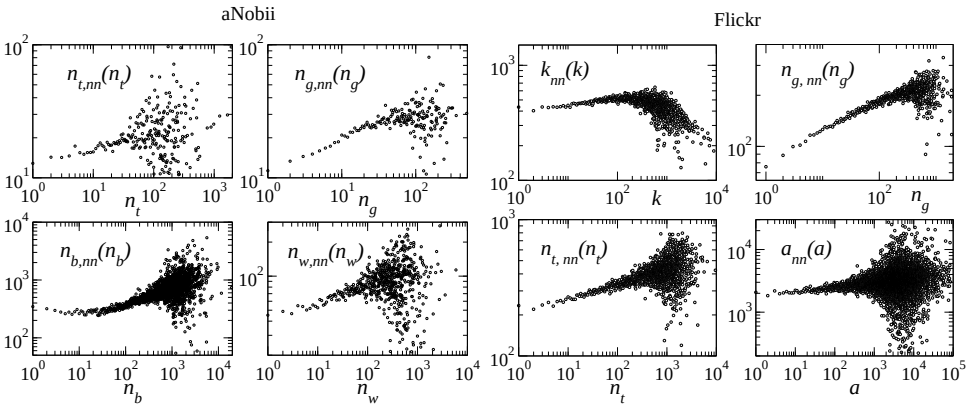

Fig. 5. Mixing patterns in the aNobii and Flickr dataset: average number $n_{t,nn}(n_t)$ of distinct tags of the nearest neighbors of users having n_t distinct tags; average number $n_{g,nn}(n_g)$ of groups of the nearest neighbors of users belonging to n_g groups; average number $n_{b,nn}(n_b)$ of books of the nearest neighbors of users who have read n_b books; average wishlist size $n_{w,nn}(n_w)$ of the nearest neighbors of users who have a wishlist of size n_w; average out-degree $k_{nn}(k)$ of the nearest neighbors of users having out-degree k; and average number $a_{nn}(a)$ of distinct triples of the nearest neighbors of users having a distinct triples.

of tags of her nearest neighbors,

$$n^u_{t,nn} = \frac{1}{k_u} \sum_{v \in \mathcal{V}(u)} n^v_t \ ,$$

and, similarly, the average total number of tags used by her nearest neighbors, $a^u_{nn} = \frac{1}{k_u} \sum_{v \in \mathcal{V}(u)} a^v$, the average number of groups to which her nearest neighbors participate, $n^u_{g,nn} = \frac{1}{k_u} \sum_{v \in \mathcal{V}(u)} n^v_g$, and, in the case of the aNobii dataset, the average number of books read by her nearest neighbors, $n^u_{b,nn} = \frac{1}{k_u} \sum_{v \in \mathcal{V}(u)} n^v_b$ and the average wishlist size of her nearest neighbors, $n^u_{w,nn} = \frac{1}{k_u} \sum_{v \in \mathcal{V}(u)} n^v_w$.

In analogy with the case of $k_{nn}(k)$, we can compute the average number of distinct tags of the nearest neighbors *for the class of users having n distinct tags*,

$$n_{t,nn}(n) = \frac{1}{|u : n_t(u) = n|} \sum_{u : n_t(u) = n} n^u_{t,nn} \ , \tag{4}$$

and the average total number of tags used by the nearest neighbors *for the class of users with a tag assignments*,

$$a_{nn}(a) = \frac{1}{|u : a(u) = a|} \sum_{u : a(u) = a} a^u_{nn} \ . \tag{5}$$

Similar formulae can be used to define the average number of groups of the nearest neighbors *for the class of users who are members of n groups*, $n_{g,nn}(n)$, the average number of books of the nearest neighbors *for the class of users who have read n books*, $n_{b,nn}(n)$ and the average wishlist size of the nearest neighbors *for the class of users who have a wishlist of size n*, $n_{w,nn}(n)$.

Figure 5 shows clear assortative trends for several measures for both the aNobii and Flickr datasets, as in other social networks [Newman and Park 2003; McPherson et al. 2001]. Similar results (not shown) are obtained for Last.fm. The average activity of the neighbors of a user increases with the user's own activity, for all the activity measures [1]. As before, large fluctuations are observed for large activity values, because of the small number of very active users. Overall, the amount of activity of socially connected users are correlated at all levels.

4.3 Topical similarity

The previous analysis has focused on the amount of user activity, as quantified by several metrics, and on the corresponding correlations and mixing patterns. To understand the interplay between the social network and user activities, it is however necessary to take into account not only the amount, but also the nature and content of the user activities. To compare users in detail, we therefore focus here on the *topical similarity* between user profiles as measured by the shared features — tags, groups, books, songs, and so on — in their profiles.

A first natural question regards the possible existence of some amount of *global* similarity between the users of a given folksonomy. For instance, in the context of tags, a simple test for the existence of a globally shared vocabulary can be performed by selecting pairs of users at random and measuring the number of tags they share, n_{st}.

In the case of Flickr, this measure shows that there is actually no shared tag vocabulary; this is not very surprising, given that Flickr is a narrow folksonomy (see Section 5) and the broad range of interests of the users. The average number of shared tags is only about 1.6 in Flickr, and the most probable case is in fact the absence of any tags shared by the selected users. When choosing two users at random this occurs with probability close to 2/3. Nonetheless, as shown in Figure 6, it can happen that randomly chosen users share a large number of tags, as the distribution of this number is quite broad and extends to values of a few hundreds tags.

Despite the lack of a globally shared profile, a number of mechanisms may however lead to *local* similarity of users' profiles, in terms of shared tags, groups membership, books, musical tastes, and so on, just as homophily effects are observed in many social networks with respect to age, ethnicity, religion, etc [McPherson et al. 2001]. The presence of a social link suggests some degree of shared context between the connected users, who are likely to have some interests in common, or to share some experiences, and who are moreover exposed to each other's content and annotations. As an example, Table II shows the 12 most frequently used tags for three Flickr users with comparable tagging activity. User A and user B have marked each other as friends, while user C has no connections to either A or B on the Flickr social network. All of these users have globally popular tags in their tag vocabulary. In this example, the neighbors A and B share an interest (expressed by the tag *flower*) and several of the most frequently used tags (marked in bold).

As often discussed in social sciences, the observed homophily can

[1]The quantitative differences between the different cases shown in Fig. 5 are not relevant to the discussion so we do not enter their detailed analysis here.

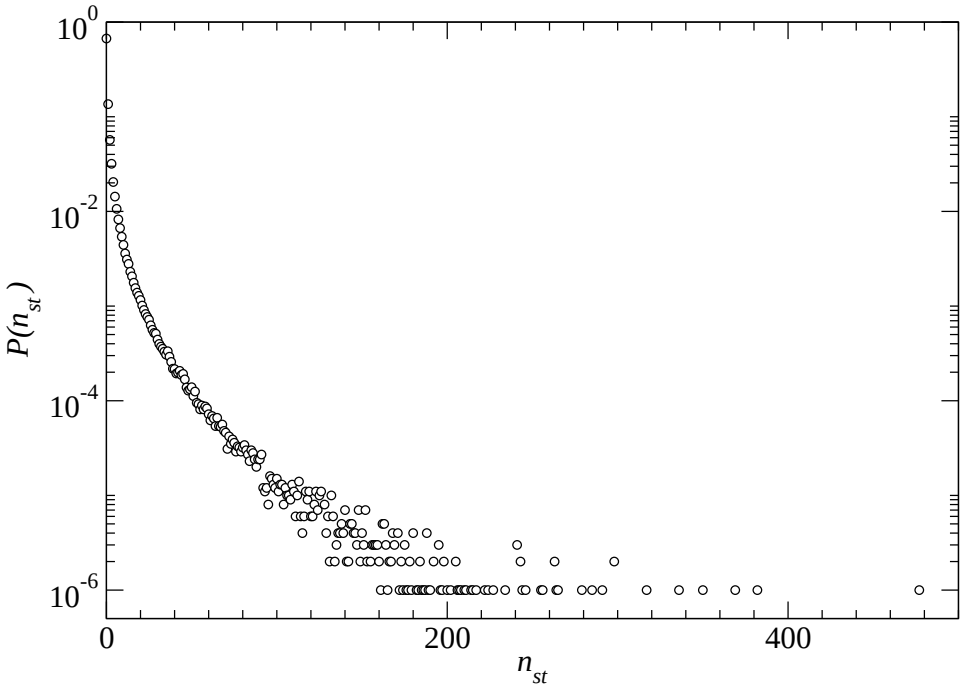

Fig. 6. Probability distribution of the number of shared tags for two randomly chosen Flickr users. The probability to have no tags in common is $P(0) \approx 0.67$, but the overall distribution is broad.

emerge for different reasons, which are summarized in two scenarios: link selection and social influence [McPherson et al. 2001; Leenders 1997; Shalizi and Thomas 2010]. The former scenario considers that social links are preferentially created between individuals who are already similar and choose each other for establishing the social link precisely because they share some degree of similarity. In the latter scenario, individuals become more similar over time because they influence each other. Disentangling these scenarios is a delicate matter that requires longitudinal data sets, as social influence implies a temporal evolution of a relationship [Crandall et al. 2008; Aral et al. 2009; Aiello et al. 2010; Shalizi and Thomas 2010]. Regardless of the distinction between these possible mechanisms driving the potential local similarity, it is important to understand how to measure this effect, and how to relate it to the social network structure, in particular with the distance between users along the network. Similarity can concern any possible type of activity: content (e.g., books in aNobii), used tags, group membership, and so on.

From this perspective, it is necessary to define robust measures of profile similarity between two users u and v, regarding the various types of activity. The first and simplest measure is given by the number of shared items for each activity: the number of shared tags n_{st} of the tag vocabularies of u and v, the number of shared groups n_{sg} to which both u and v belong, the number of common books in

Table II. Tags most frequently used by three Flickr users

User A	User B	User C
green	**flower**	japan
red	**green**	tokyo
catchycolors	kitchen	architecture
flower	**red**	bw
blue	**blue**	setagaya
yellow	white	reject
catchcolors	fave	sunset
travel	detail	subway
london	closeupfilter	steel
pink	metal	geometry
orange	**yellow**	foundart
macro	zoo	canvas

their libraries or wishlists for aNobii, and the number of common songs for Last.fm. These measures may however be affected by the amounts of activity of the users; two users who apply many tags may have more tags in common than two less active users, just because it is more probable to find common items in two long lists than in two short ones. For instance, let us consider two users with 100 tags each, and having 10 of them in common. The number of shared tags is 10 in this case, but represents just 10% of their tagging activity. Two users with the same 5 tags, on the other hand, have $n_{st} = 5$, i.e. less than in the previous case, but this represents 100% of their activity. In short, such simple measures are not normalized, and we therefore also need to consider measures that compensate for the heterogeneity in the amounts of activity. To this end, we consider a distributional notion of similarity between the profiles of u and v.

Let us first consider the case of the tags. Following Cattuto et al. [2008] we regard the vocabulary of a user u as a *feature vector* W whose elements correspond to tags and whose entries are the tag frequencies for that specific user's vocabulary, i.e., w_{ut} is the number of resources tagged with t by u. To compare the tag feature vectors of two users, we use the standard cosine similarity [Salton 1989] defined as

$$\sigma_{tags}(u,v) = \frac{\sum_t w_{ut} w_{vt}}{\sqrt{\sum_t w_{ut}^2}\sqrt{\sum_t w_{vt}^2}} \,. \tag{6}$$

This quantity is 0 if u and v have no shared tags, and 1 if they have used exactly the same tags, in the same relative proportions. Because of the normalization factors in the denominator, $\sigma_{tags}(u,v)$ is not directly influenced by the global activity of a user.

Similarly, we can define the cosine similarities for groups memberships and for books. Since a user belongs at most once to a group, and adds a book only once to her library, the elements of the group and book vectors are binary, and the cosine similarity reduces to

$$\sigma_{groups}(u,v) = \frac{\sum_g w_{ug} w_{vg}}{\sqrt{n_g(u) n_g(v)}}; \quad \sigma_{books}(u,v) = \frac{\sum_b w_{ub} w_{vb}}{\sqrt{n_b(u) n_b(v)}}, \tag{7}$$

where w_{ug} is 1 if u belongs to group g and 0 otherwise, and w_{ub} is 1 if u has book b in her library and 0 otherwise.

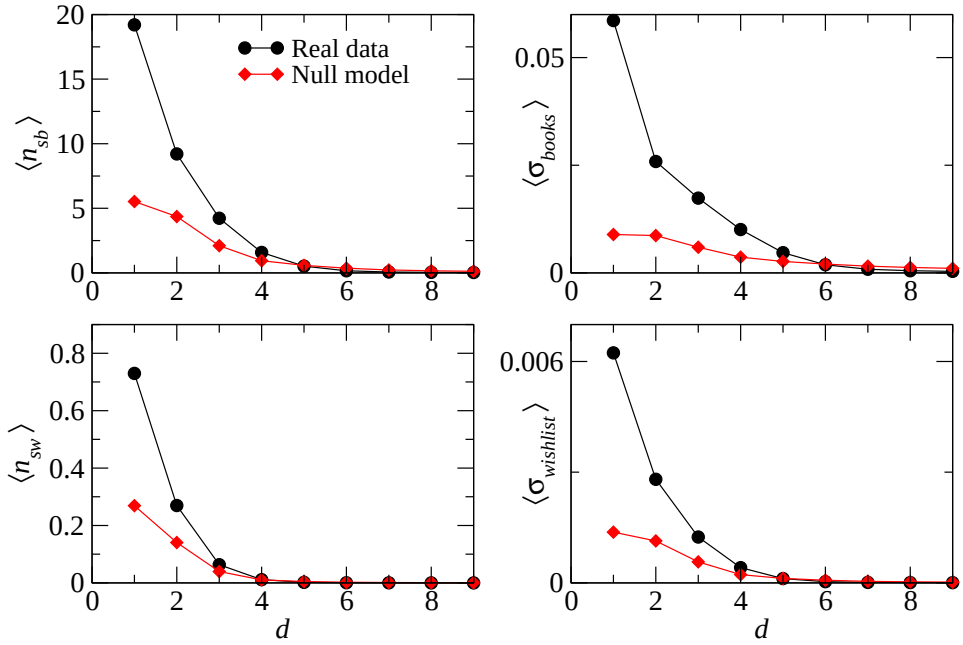

Fig. 7. Average library and wishlist similarity as a function of the distance on the aNobii social network. Top: average number of shared books in the libraries of users at distance d, and corresponding cosine similarity. Bottom: average number of shared items in the wishlists of users at distance d, and corresponding cosine similarity. The diamonds correspond to the null model discussed at the end of Section 4.

Figures 7 and 8 give an indication of how the similarity between users depends on their shortest path distance d on the social network, by showing the average similarity of two users as a function of d. In aNobii, for instance, the average number of shared books is rather large for neighbors (close to 20), but it drops rapidly as d increases, and is close to 0 for $d \geq 4$. Similar results are obtained for the number of common groups and tags, and hold for Last.fm and Flickr as well. The cosine similarities display the same decreasing trend as the distance along the social network increases.

The shortest path distance between two users gives the minimum number of steps to navigate on the online social network to go from one user to the other. This measure of topological proximity between users can however be sensitive to the addition or removal of one single link, and does not take into account the fact that more than one path can connect the users. To overcome this issue, the personalized PageRank [Haveliwala 2003] of one user v with respect to another user u can be considered. This personalized PageRank essentially gives the probability, for a random walker starting from the profile page of user u, to visit the profile page of v. As shown in Fig. 9, the topical similarity between users increases when their relative personalized PageRank increases. As the personalized PageRank decreases when the distance between users

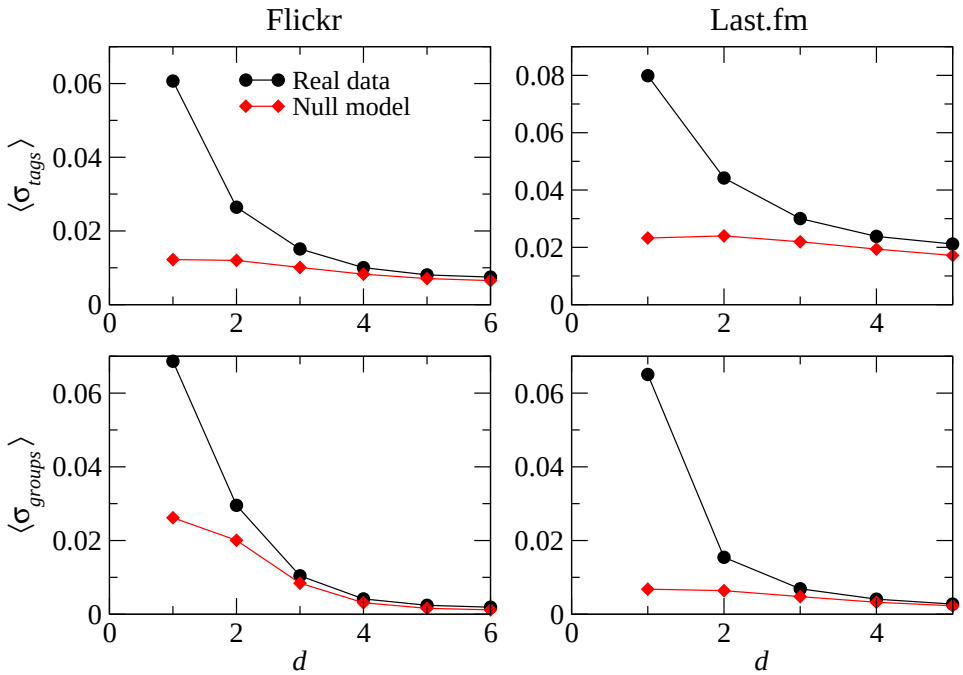

Fig. 8. Average tag and group similarity as a function of the distance on the Flickr and Last.fm social networks. Top: average cosine similarity between the tag vocabularies. Bottom: average cosine similarity between the groups participation vectors. The diamonds correspond to the null model discussed at the end of Section 4.3.

increases, this increase is consistent with the decreasing trend of Fig. 7.

To gain more insight into the entanglement between similarities and distance on the social network, we present in Figures 10 and 11 the probability distributions of the selected similarity measures for pairs of users at social distance d. The figures clearly expose the dependence of all distributions upon the distance of the users along the social network: for users who lie at small distances on the social network, rather broad distributions spanning several orders of magnitude are observed for the number of shared tags, groups, or books. As the distance d along the network increases, the distributions become narrower. Two comments are in order: first, the distributions of n_{st} at short distances reach much larger values of n_{st} than in Figure 6 (the same is observed for the number of shared groups or books). The reason is that, when choosing a random pair of nodes (as in Figure 6), it is very unlikely to select two neighboring nodes. Second, at any distance, the most probable value of n_{sg} or n_{st} is 0, even if the distributions are broad, and this probability increases with d. For instance, for Flickr users, the probability $P(n_{st} = 0)$ that two users do not share any tag is 0.1 if the users are neighbours (i.e., at $d = 1$), 0.17 if they are at distance $d = 2$, 0.37 at distance $d = 3$. For groups, we obtain $P(n_{sg} = 0)$ is 0.17 for $d = 1$, 0.4 at $d = 2$, 0.74 at $d = 3$.

The distributions of cosine similarities between users at distance d show similar

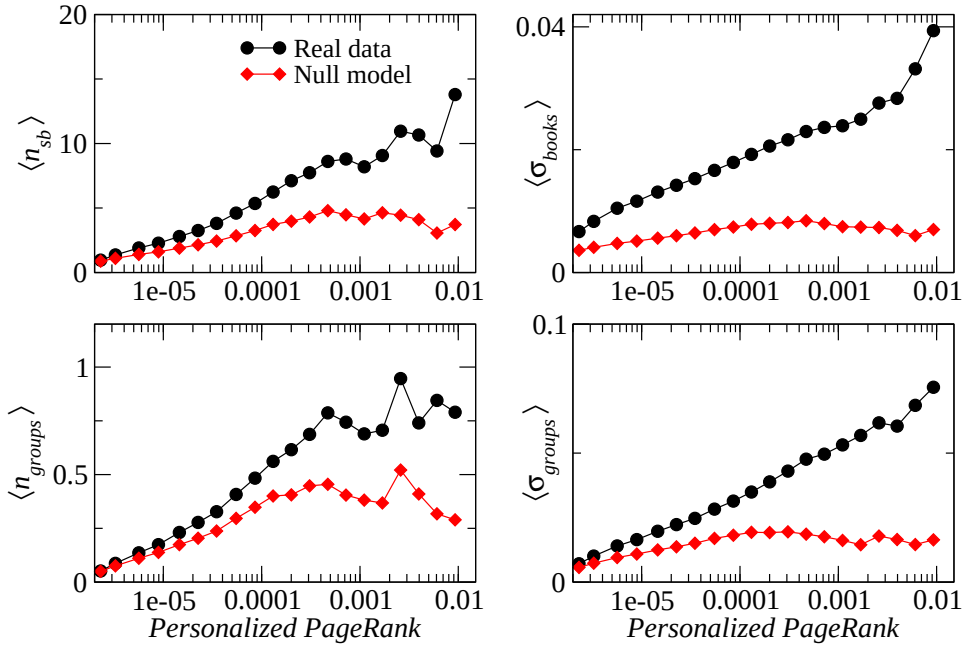

Fig. 9. Average library (top) and group (bottom) similarity between two aNobii users as a function of the personalized PageRank of one user with respect to the other user. The diamonds correspond to the null model discussed at the end of Section 4.

features: they mostly span the whole interval of possible values, but the probability of reaching values close to 1 becomes smaller when d increases, while the probability of a zero similarity increases. The trends are similar for Last.fm.

The presence of assortative mixing patterns in the social network, with respect to the intensity of users activity, makes it necessary to investigate in more detail the observed local similarity of profiles. It could indeed be the case that such assortativity, by a purely statistical effect, yields an *apparent* local similarity between the tag vocabularies of users. For example, even in a hypothetical case of purely random tag assignments, it would seem more probable to find tags in common between two large tag vocabularies than between a small one and a large one. Furthermore, as we have shown, users who are more active have more friends, and their friends are also more active, therefore similarity with their friends may depend on their greater activity alone.

To discriminate between effects simply due to the assortativity and those due to actual profile similarity, one has to construct a proper *null model*, i.e., an artificial system that retains the same social structure as the one under study, but lacks any feature similarity other than the one that may result from purely statistical effects. This is done by keeping fixed the social network and its assortativity pattern for the intensity of the activity, but destroying socially-related feature similarity by means of a random permutation of profile items. For instance, we proceed in the following fashion for the tags: (i) we keep the social network unchanged, preserving each user's degree k; (ii) we shuffle the tags among users in such a way as to preserve

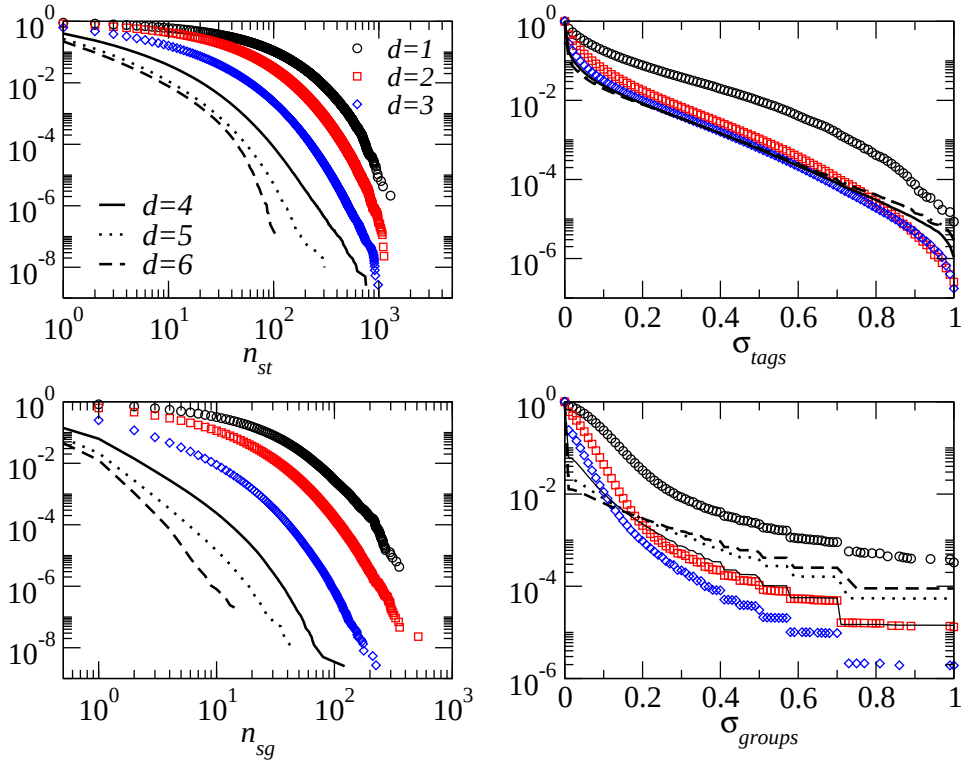

Fig. 10. Left: Complementary cumulative probability distributions of the number of shared tags and groups for two Flickr users lying at distance d on the social network, for different values of d. Right: Complementary cumulative probability distributions of the cosine similarity between the tag vocabularies and group memberships of two Flickr users.

each user's number of tag assignments a as well as number of distinct tags n_t. This guarantees that the distribution of frequencies of tags is left unchanged. For group membership and for books, we can proceed in a similar way except we preserve in the shuffle each user's number of groups n_g and number of books n_b. **This procedure is in the spirit of the definition of null models for detecting the importance of patterns in networks [Maslov et al. 2004], or of the definition of random models of networks with given degree distributions or correlation patterns [Molloy and Reed 1995; Catanzaro et al. 2005; Serrano and Boguñá 2005]**.

This null model preserves the assortativity patterns with respect to the amount of user activity, as each user has exactly the same number of distinct tags, tag assignments, groups, and books as in the real data. However, correlations between the tag vocabularies and other features are lost, except for the ones purely ascribed to statistical effects.

Using the null model defined above, we measure the similarity between users at distance d on the social network in the same way as for the original data. As Figure 7 shows, the average number of books in libraries and wishlists, as a function of the

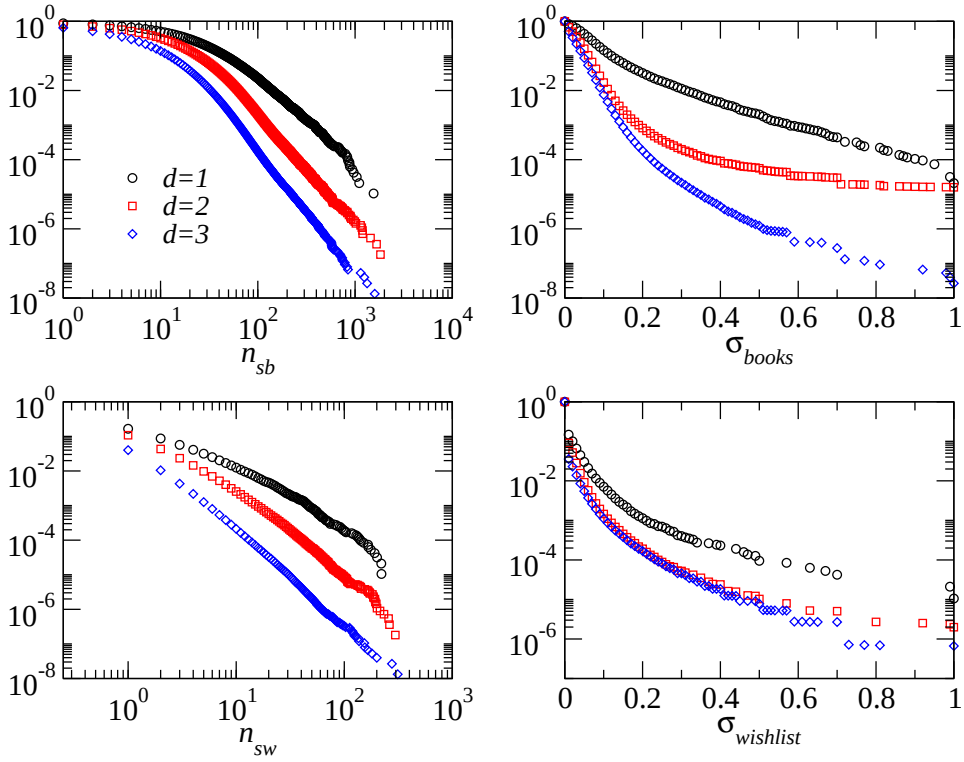

Fig. 11. Complementary cumulative distributions of the number of shared books and (left), and of the similarities in the lists of books (right), in the libraries (top) and in the wishlists (bottom) of aNobii users lying at distance d on the social network, for various values of d.

distance d, shows a very similar trend to the original (non-shuffled) data. Similar curves are obtained for the number of shared tags or groups. For neighboring users, and also for next-to-nearest neighbors, the average numbers of shared tags or groups are generally significantly lower in the null model, but the distributions are very similar, as shown in Figure 12(top). The assortative mixing between the amount of activity of neighboring users is therefore enough to yield a strong topical similarity *as simply measured by the number of shared tags, groups or books*. The case of the cosine similarity is quite different: as shown in Figures 7 and 8, the average cosine similarity in the null model does not depend as strongly on distance in the social network. Figure 12(bottom) also shows that the distributions of σ_{tags} are very different for the original and shuffled data, and do not depend on distance in the case of the shuffled data.

We conclude that the homophily measured by the cosine similarity is a genuine non-random effect in these social networks, not only due to the assortative mixing: the measured topical homophily is not only due to the homophily in terms of amounts of activity.

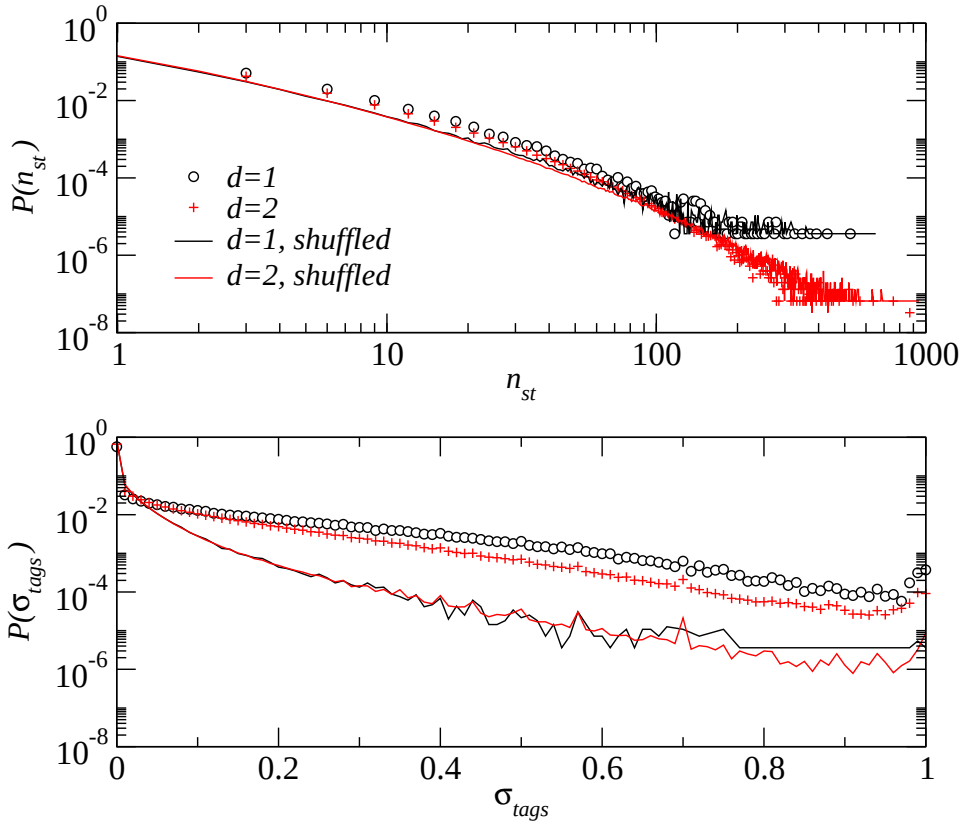

Fig. 12. Top: Probability distributions of the number of shared tags of two Last.fm users lying at distance d on the social network, for $d = 1$ and $d = 2$ (symbols), and for the same network with shuffled tags (lines). Bottom: same for the distributions of the cosine similarities of the tag vocabularies.

5. SOCIAL LINK PREDICTION

In the previous section, we have observed the presence of homophily in the social network: users connected by a social link show a significantly higher topical similarity compared to non-linked users. This correlation is observed for several activities, like tagging and group affiliation, and occurs in all the datasets considered. A consequent guess resulting from such finding is that the presence of a social tie could be predicted relying only on the topical similarity between users. Since we obtained information about user activity and social links for all the datasets discussed above, we are able to test this hypothesis in three different settings.

In this section, we focus on the Last.fm and aNobii datasets. We do so for several reasons. First, in these two cases, the data include information on the users' libraries, in addition to groups and annotations. Second, a prediction based on the tagging information is more meaningful in a *broad* folksonomy, in which users can label the same (global) set of items and pick from this global set to form their libraries. This condition allows us to deal with similarity based on shared

content as well as shared vocabulary. Last.fm and aNobii are broad folksonomies as any user can tag any artist or book. This is not the case for Flickr, considered a narrow folksonomy. Users usually tag only the pictures they upload themselves.

Moreover, the Last.fm and aNobii datasets have peculiarities that allow to draw interesting conclusions regarding link prediction task. They both have a detailed specification about the users' mother tongue. As we will see, language is a feature that can considerably influence the prediction, and that should thus be taken into account when accuracy is measured. Additionally, Last.fm provides the *tasteometer* score, a user-to-user similarity metric computed by the system. **The tasteometer algorithm is not public**. Since tasteometer is independent from any social information (as we have verified empirically, although we do not have a precise information of how the tasteometer values are effectively computed) and is based only on listening patterns, we can fairly compare the prediction accuracy achieved by user-to-user topical similarity measures and the one obtained by the system-provided similarity metric.

5.1 Methodology

Our link prediction problem can be defined as follows. Given a subset of users $U_t \subseteq U$, we want to predict the presence or absence of a social link for every pair $(u, v) \in \{U_t \times U_t | u \neq v\}$. Of course, the information about the social network topology is not known, but we are given the full information about the *features* that describe the user profiles. We deal in particular with four different features: *groups*, *library*, *tags*, and tagged *items*. Note the difference between *items* and *library* features. In aNobii, tagged items are a subset of the whole set of books in the user's library, while in Last.fm tagged items can be tracks, albums or artists and the library is composed only by the top 50 artists in the user's global playlist. Tags and items can be directly extracted from the three-dimensional triples space through aggregation (details are provided in Section 5.2).

We take into account each feature separately, so each user is described with a single *feature vector*. For each pair of users in U_t, we compute a similarity value between their feature vectors using the metrics defined in Section 5.2. In the case of Last.fm we also have the system-provided tasteometer similarity. Next, we sort the node pairs in decreasing order of their similarity score. The pairs with the highest topical similarity are those that we suppose are the most likely to be connected with a social link. For this reason, we predict the presence of a tie for every user pair whose similarity value is greater than or equal to a threshold value σ. To evaluate the accuracy of our predictions, we check the presence of each predicted link in the real social network and we count the number of true positives and false positives. As the value of σ decreases, a higher number of links is predicted, leading to an increase in the number of both true positives and false positives. We test the accuracy of our predictor for all the significant values of σ. The similarity measure that performs best for the prediction task is the one that achieves the best ratio between true positives and false positives, across all the possible threshold values. To quantitatively measure the prediction performance for the whole set of threshold values we consider ROC curves [**Fawcett 2006**] and we compare the area under the curve (AUC) achieved by the different features and similarity metrics considered. **ROC curves are commonly used in the machine learning community for**

the link prediction task [Clauset et al. 2008].

Given this setting, it is important to select a significant sample of users U_t. Intuitively, one could choose $U_t \equiv U$. The problem is that, since the social graph's density is very low, the full social matrix $U \times U$ is very sparse, thus leading to a very low number of potential true positives.

The problem of studying greatly biased datasets is well-known by data miners and it is a common issue also in social link prediction due to the intrinsic sparsity of social graphs [Liben-Nowell and Kleinberg 2003; Getoor and Diehl 2005; ?; Lü and Zhou 2010]. It has been shown that the AUC is a good measure for the performance evaluation when there is a strong class skew [Stäger et al. 2006]. However, in the first part of our evaluation we want to minimize the sparsity problem in order to compare the predictive power of different features in a less biased setting.

We thus restrict our analysis to several smaller subsets each composed of 500 users only, sampled on the basis of one of two criteria. First, we extracted a *Most Connected* set for each feature, composed of the nodes with the highest out-degree and that have at least one element for the considered feature. Second, we sampled a distinct *Most Active* set for each feature, containing the 500 users with the largest number of elements for that feature. More in detail, we chose the sets of users with the highest number of groups, with the highest number of objects in their libraries, and, for both item and tag features we chose the set of the 500 taggers with the highest number of triples. In short, we have distinct Most Active and Most Connected samples for each feature, except for items and tags because they both derive from the same triple space.

The Most Active sampling provides the best scenario in which to explore the effectiveness of link prediction based on topical similarity Furthermore, given the correlation between user activity and social connectivity (see Figure 3), the Most Active nodes typically have a rather high degree, thus ensuring a relevant number of intra-sample social connections. As a result, the density of our social network samples ranges from 0.02 to 0.07, which is three–four orders of magnitude higher than the full networks.

In a second phase of the evaluation, we expand our observations with a sensitivity analysis to show how much the prediction accuracy is affected by the sub-graph density, the user activity, and the sample size, thus disentangling the evaluation from the possible skew due to the narrower selection of the most active and connected users.

5.2 Similarity metrics

To model the task of predicting social links we need to define measures of profile similarity between users. In particular, we have to select a robust similarity metric for the features that characterize the activity of users. In relation to the groups membership and library features we follow the approach in Section 4.3 that computes similarities by way of the standard cosine similarity as formalized in Equation 7.

For the remaining features we adopt the framework by Markines et al. [2009] that represents the system as a *tripartite* graph that involves users, tags, and resources (e.g., books, songs, photos, etc.). A ternary relation between a user u, a tag t, and

a resource r can be defined as a *triple*. We then can establish a set of triples as a folksonomy F and we can define similarity measures $\sigma(x, y)$ where x and y can be two resources, tags, or users. In our analysis we focus on the definition of similarity functions $\sigma(u, v)$ where u and v are two users.

Since measures for similarity and relatedness are not well developed for three-mode data such as folksonomies, we consider various ways to obtain two-mode views of the data. In particular, we consider two-mode views in which the two dimensions considered are dual — for example, users can be represented as sets of tags or resources. The process of obtaining a two-mode view from a folksonomy is called *aggregation*. In the remainder of this section, we will discuss different aggregation strategies and the set of similarity functions we adopted.

5.2.1 *Aggregation Methods.* In reducing the dimensionality of the triple space, we necessarily lose correlation information. Therefore, the aggregation method is critical for the design of effective similarity measures; poor aggregation choices may negatively affect the quality of the similarity by discarding informative correlations. As mentioned above, we can define similarity measures for each of the three dimensions (users, resources, tags) by first aggregating across one of the other dimensions to obtain a two-mode view of the annotation information. Since our analysis is intended to explore user similarity we can aggregate across one of the tag or resource dimensions, obtaining a description of a user as a vector of, respectively, resources or tags.

We will consider four approaches to aggregate user information: *projection*, *distributional*, *macro*, and *collaborative* aggregation. To simplify our exposition, in the following definitions we will adopt an aggregation across resources, meaning that a user will be represented as a vector of tags; analogous mechanisms apply when tags are selected as the aggregation dimension. An extensive discussion on these aggregation approaches can be found in our prior work [Markines et al. 2009].

Projection. The simplest aggregation approach corresponds to the projection operator $\pi_{u,t}(F)$ in relational algebra, assuming the triples are stored in a database relation F. Another way to represent the result of aggregation by simple projection is a matrix with binary elements where rows correspond to users (as binary vectors, or sets of tags) and columns corresponds to tags (as binary vectors, or sets of users).

Distributional. A more sophisticated form of aggregation stems from considering distributional information associated with the set membership relationships. One way to achieve distributional aggregation is to make set membership fuzzy, i.e., weighted by the Shannon information (log-odds) extracted from the annotations. Intuitively, a tag shared by two users may signal a weak association if it is very common. For example, let U be the set of users and U_t the users that annotate with t. We will use the information of tag t defined as $-\log p(t)$ where

$$p(t) = \frac{|U_t|}{|U|}. \tag{8}$$

Another approach is to define a set of *frequency-weighted* pairs (u, t, w_{ut}) where the weight w_{ut} is the number of resources tagged with t by u. Such a representation corresponds to a matrix with integer elements w_{ut}, where rows are user vectors and

columns are tag vectors. We will use both of the above distributional aggregation approaches as appropriate for different similarity measures.

Macro. **To compute an average function in class-partitioned datasets (e.g., documents partitioned in categories), micro- and macro-averaging approaches are possible. Micro-averaged scores are calculated considering the contribution from each element in each class. In contrast, macro-averaged values are obtained by first calculating the function for each class and then taking the average of the results. Micro-averaging gives equal weight to every element while macro-averaging gives equal weight to every class. Both approaches are broadly used in text mining [Feldman and Sanger 2006]**. By analogy, distributional aggregation can be viewed as *micro-aggregation* if we think of resources as classes. Each annotation is given the same weight, so that a more popular resource would have a larger impact on the weights and consequently on any derived similarity measure. In contrast, macro-aggregation treats each resource's annotation set independently first, and then aggregates across resources. This will allow the similarity calculation to be *incremental*, breaking the dependency on global frequencies. In relational terms, we can select the triples involving each resource r, and then project, yielding a set of pairs for r: $\{(u,t)_r\} = \pi_{u,t}(\sigma_r(F))$. This results in per-user binary matrices of the form $w_{r,ut}$. The per-user binary matrix representations $w_{r,ut} \in \{0,1\}$ are used to compute a local similarity $\sigma_r(u,v)$ for each pair of users u and v. When defining the Shannon information of a feature, the feature probability $p(t)$ must be replaced by a conditional probability $p(t|r)$. Finally, we macro-aggregate by voting, i.e., by summing across resources to obtain the global similarity. Macro-aggregation does not have a bias toward resources with many annotations. However, in giving the same importance to each resource, the derived similarity measures amplify the relative impact of annotations by less popular resources.

Collaborative. Macro-aggregation lends itself to the exploration of collaborative filtering in folksonomies while the computation remains incremental. Thus far, we have only considered feature-based representations when working with a tripartite representation. That is, a user is described in terms of its tag or resource features. If two users share no feature, all of the measures defined on the basis of the aggregation schemes will yield a zero similarity. In collaborative filtering, on the other hand, the fact that one or more users vote for (or in our case annotate) two objects is seen as implicit evidence of an association between the two objects, even if they share no features. The more users share a pair of items, the stronger is the association. We want to consider the same idea in the context of user similarity in folksonomies. If many resources have been annotated by the same pair of users, even with different tags, the two users might be related. Likewise, if two users apply the same tags, even to annotate different resources, the two users might be related. We can capture this by adding a feature-independent local similarity to every pair (u,v) of users in macro-aggregation. In practice we can achieve this by adding a special "resource tag" t_r to all users that tagged r. This way all of r's users have at least one annotation in common. However, the information of such special tag would be $-\log(p(t_r|r)) = -\log(1) = 0$. To ensure that the special tag makes a non-zero contribution to the local similarity $\sigma_r(u,v)$, let us redefine the odds of tag t for

resource r as

$$p(t|r) = \frac{|u_{t,r}|}{|u_r| + 1} \tag{9}$$

which is always less than 1 so that $-\log(p(t_r|r)) > 0$.

5.2.2 *Similarity Measures.* We wish to explore several information-theoretic, statistical, and practical similarity measures. Each of the aggregation methods requires revisions and extensions of the definitions for application to the folksonomy context. Having shown in our prior work [Markines et al. 2009] that distributional aggregation yields better accuracy than projection, and collaborative aggregation yields better accuracy than macro-aggregation, with the same computational complexity, we focus on distributional and collaborative aggregations. For brevity, we show definitions only for the similarity measures in the distributional case. These definitions are based on feature probabilities $p(x)$ defined in Equation 8. The definitions of the local similarities for collaborative aggregation are similar except that the feature probabilities are replaced by the conditional probabilities defined in Equation 9. We suppose that $u, v \in U$ represent users and X_u, X_v are their vector representations. Of course, the attributes of X depend on the aggregation dimension. In the following formulas we consider the case of aggregation across resources, i.e. the users are denoted by vectors of tags with tag elements w_{ut}.

Matching. The distributional version of the matching similarity is

$$\sigma(u, v) = -\sum_{t \in X_u \cap X_v} \log p(t). \tag{10}$$

Overlap: Distributional overlap is given by

$$\sigma(u, v) = \frac{\sum_{t \in X_u \cap X_v} \log p(t)}{\max(\sum_{t \in X_u} \log p(t), \sum_{t \in X_v} \log p(t))}. \tag{11}$$

Jaccard: Distributional Jaccard similarity is defined as

$$\sigma(u, v) = \frac{\sum_{t \in X_u \cap X_v} \log p(t)}{\sum_{t \in X_u \cup X_v} \log p(t)}. \tag{12}$$

Dice: Distributional version of Dice is defined as

$$\sigma(u, v) = \frac{2 \sum_{t \in X_u \cap X_v} \log p(t)}{\sum_{t \in X_u} \log p(t) + \sum_{t \in X_v} \log p(t)}. \tag{13}$$

Cosine: For the distributional version of the cosine, it is natural to use the frequency-weighted representation

$$\sigma(u, v) = \frac{X_u}{||X_u||} \cdot \frac{X_v}{||X_v||} = \frac{\sum_t w_{ut} w_{vt}}{\sqrt{\sum_t w_{ut}^2} \sqrt{\sum_t w_{vt}^2}}. \tag{14}$$

This formula is equivalent to Equation 6.

Maximum Information Path: The last measure we consider is *Maximum Information Path* (MIP) [Markines and Menczer 2009]. The MIP similarity is an extension of traditional shortest-path based similarity measures and Lin's similarity measure [Lin 1998]. MIP differs from traditional shortest-path similarity measures

by taking into account Shannon's information content of shared tags (or resources). Lin's similarity measure only applies to hierarchical taxonomies, such as the case when bookmarks are organized in folders and subfolders. However, when the folksonomy includes non-hierarchical annotations, Lin's measure is undefined while MIP similarity is well defined and captures the same intuition. The association between two objects is determined by the ratio between the maximum information they have in common (most informative shared feature) and the information they do not share. Because of the dependency on log-odds, maximum information is not defined for projection aggregation. We define MIP for the distributional case as

$$\sigma(u,v) = \frac{2\log(\min_{t\in X_u\cap X_v}[p(t)])}{\log(\min_{t\in X_u}[p(t)]) + \log(\min_{t\in X_v}[p(t)])}. \tag{15}$$

In prior work we have also explored mutual information and found it to be competitive but expensive [Markines et al. 2009], therefore we exclude it from the present analysis.

5.2.3 *Performance evaluation.* **When computing similarity in large systems, the issue of scalability becomes crucial. The most important factor which affects scalability in the computation of the similarity matrix is the aggregation method adopted. In the following, we perform a computational complexity analysis focusing on distributional and collaborative aggregations, since they proved to be far more effective than other known aggregations [Markines et al. 2008].**

Formally, both aggregations lead to a $O(N^2)$ complexity, where N is the size of the system in terms of number of users; however, the major difference between the two approaches is *incrementality*. With distributional aggregation, similarities must be recomputed from scratch whenever new triples are added to the system, as frequency weights must be updated. Conversely, collaborative aggregation allows for incremental computation because each new triple affects only the contribution that the incoming tag or resource (depending on the aggregation dimension) gives to the overall similarity matrix.

From a practical point of view, we can consider as scalable those measures that can be updated as a stream of incoming annotations is received. However, since the update time clearly depends on how many user pairs' similarity scores are affected by the new triples, we should study how the update time changes as the system size grows.

We recur to an empirical analysis to examine how the update complexity scales with the number of users and triples in the system. Figure 13 shows the complexity for the two different flavors of collaborative aggregations and a single representative case for distributional aggregation, since in this case the aggregation dimension does not impact on performance. The experiment is performed on the aNobii dataset, using the MIP similarity.

Curves clearly show that collaborative aggregation over items outperforms aggregation over tags. This result is basically due to the different distributions of items and tags over users. When a new triple is added,

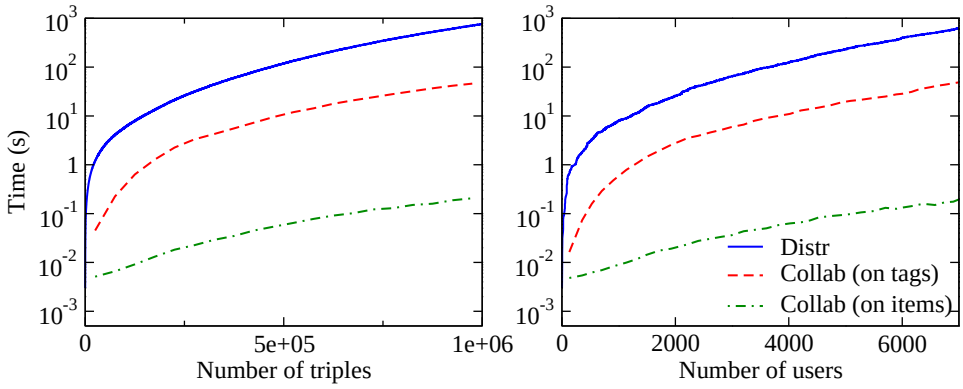

Fig. 13. Scalability of the MIP similarity computation for distributional and collaborative aggregations. Scalability of collaborative aggregation is shown for two different aggregation dimensions (items and tags). The figure reports the CPU time (in seconds) against the number of triples and users in the system.

the contribution of the value of the aggregation dimension of that triple to the overall similarity matrix should be recomputed. The greater is the number of users who have that value in their triple sets, the higher is the number of pairs whose similarity score must be updated. In the considered sample the number of resources in the triple set is one order of magnitude larger than the number of distinct tags; similar ratios hold in any big folksonomy. For this reason, it is more likely that the addition of a triple containing a very popular tag affects many more users (and consequently the similarity between them and others) than a triple with a very popular item.

However, note that collaborative aggregation takes a little time to update a system with a great number of triples even if recurring to aggregation over items.

5.3 Prediction with single features

We computed the similarity metrics on the Most Active and Most Connected sets of users for the four features considered, on the aNobii and Last.fm datasets. For the two folksonomy-related features, items and tags, we combined all the aggregation methods with all the similarity metrics defined above (except for the projection-MIP combination, which is not defined). For groups and library features we calculated the similarity using matching, overlap, Dice, Jaccard, and cosine metrics; note that these features do not require any aggregation and MIP is not defined.

We consider also two additional metrics as baselines. First, we queried the Last.fm API service for the tasteometer scores related to the same Most Active and Most Connected samples used for items and tags. **Second, we computed the similarity in terms of number of common neighbors (CN) for the Most Connected samples which, in this case, overlaps with the Most Active sample since the number of connections is the feature considered. We introduce this widely used metric to compare the performance**

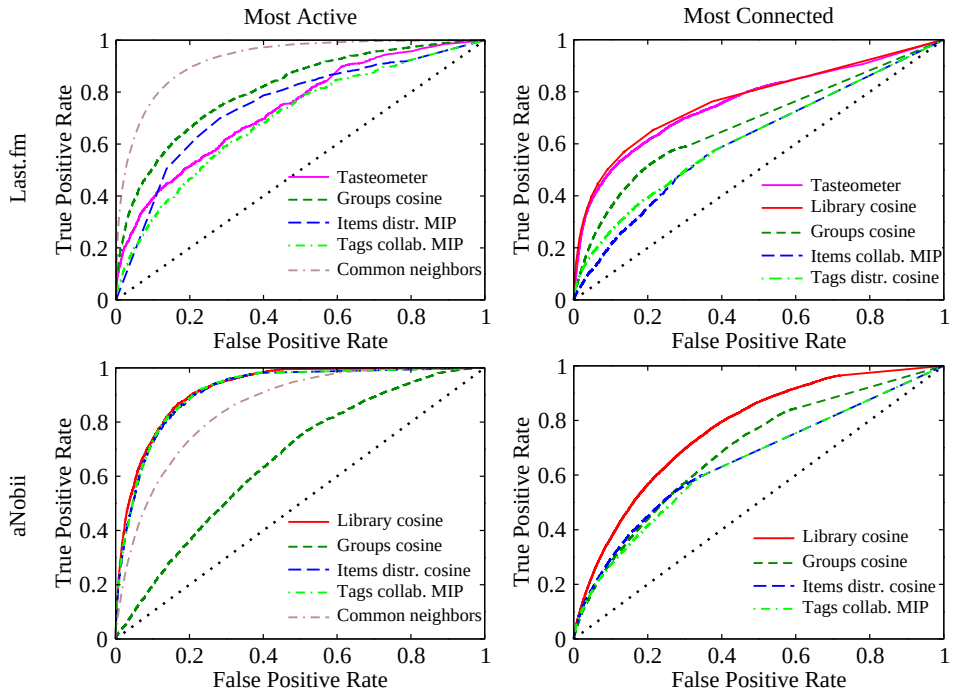

Fig. 14. ROC curves comparing the accuracy for the best feature-based predictions in both datasets. The tasteometer baseline curve is also shown for Last.fm.

of network-based and feature-based similarities. Among all the known network-based metrics we opted for CN because, despite its simplicity, it has been shown to be an effective predictor of social ties and because it is a local measure, whose computation is scalable. Altogether we obtained 133 **user-to-user similarity networks that we used to perform as many social link predictions.**

For brevity, we next report only on a selection of representative cases, restricting our evaluation to the best-performing instances. In particular, when we deal with items and tags, we refer to the cosine and MIP similarity, computed using both distributional and collaborative aggregation. We choose the cosine similarity as a representative case also for the groups and library features. AUC values for the Last.fm and aNobii networks are summarized in Table III. Note that since the Last.fm library provided via the API has a size bounded to 50 artists, we cannot identify the Most Active users for this feature, therefore we omit the Most Active sample for the library feature.

Not surprisingly, most of the the highest AUC values are achieved for the Most Active samples, because of the greater amount of information available.

Regarding the folksonomy-based features, we find that the MIP similarity often outperforms the cosine metric, as well as the other measures (not shown). Furthermore, for the Most Active Last.fm users represented through items, MIP similarity outperforms the tasteometer baseline. Another important result emerging from the

Table III. AUC values for Last.fm and aNobii social link predictions calculated for the four user features. The user samples considered are the most active with reference to the considered feature and the most connected users that have at least one element for that feature. The Last.fm results refer to one of our three snapshots; results for the other snapshots are consistent. The tasteometer similarity was calculated for the same Most Active and Most Connected sets used for items and tags. The feature vectors for items and tags have been obtained through Distributional or Collaborative aggregation over the three-dimensional folksonomy. Shown in bold are the best results for each combination of dataset, sampling method and feature.

		Last.fm		aNobii	
Feature	Similarity	Active	Connected	Active	Connected
Baselines	Tasteometer	0.734	0.759	-	-
	Common neighbors	0.927	-	0.854	-
Items	Distrib cosine	0.663	0.560	**0.915**	**0.655**
	Distrib MIP	**0.749**	0.559	0.878	0.649
	Collab MIP	0.589	**0.613**	0.652	0.561
Tags	Distrib cosine	0.579	**0.625**	0.652	0.554
	Distrib MIP	0.697	0.618	0.651	0.560
	Collab MIP	**0.698**	0.559	**0.916**	**0.648**
Groups	Cosine	**0.810**	**0.677**	**0.662**	**0.690**
Library	Cosine	-	**0.769**	**0.923**	**0.768**

data is that the aggregation process has a great impact on the predictive potential; we find that when describing users through item vectors (aggregating across tags) the distributional approach tends to be more profitable, while when representing users as vectors of tags the collaborative aggregation tends to be preferable.

Predictions based on groups and libraries perform even better. For groups, we note a lower accuracy in the aNobii case compared to Last.fm, due to the relatively low cardinality of the group set; in fact, in aNobii we have about $3,000$ groups, against about $70,000$ groups in Last.fm. Inevitably, a lesser range of choice corresponds to a greater uniformity in the group affiliation behavior, thus making it more difficult to infer social connections. Lastly, the library results suggest that this is the best feature for prediction purposes, when applicable. Since in Last.fm the library feature vectors have at most just 50 elements each, the implicit social information carried by the elements in the library is very high. When the cardinality of the feature vector is unbounded, like in the Most Active scenario in aNobii, the AUC values become even higher. Of course, for aNobii, we expected the library-based prediction to be more accurate than the item-based ones, simply because the set of books in the library is a superset of the tagged books that can be retrieved from the folksonomy. Nevertheless, we observed that peak AUC values in aNobii are in part determined by the particularly strong geographically-biased clustering of its social network, which we discuss in detail in Section 6. **Finally, we observe that the prediction based on common neighbors can be very accurate when a lot of information about social contacts is available. In Last.fm's Top Active sample, CN is even the best performing metric, while in aNobii the library and the folksonomic features are more accurate.**

In Figure 14 we depict a summary comparison between the ROC curves of the best performing prediction measures, i.e., those shown in bold in Table III.

The analysis on small user samples is useful to compare the effectiveness of different metrics under different boundary conditions of connec-

tivity and activity. However, to show that the results are not biased by this sampling procedure, we performed a *sensitivity analysis* to see how the AUC value changes when connectivity, activity and sample size are varied. To do so, first we sort users in decreasing order of activity and connectivity. Then we collect samples with decreasing activity and connectivity by placing on each list a sliding window with a size of 500 and shifting it by 250 users at time. To collect samples with different sizes we simply compute the AUCs for the top K elements of the list, with K going up to $5,000$.

Results for the library feature in aNobii are depicted in Figure 15; the same qualitative results hold for other features. Together with AUC, we measure also the precision at top N i.e, how many of the pairs which have the N highest similarity values are actually connected; this is a metric which is commonly used to complement the AUC [Backstrom and Leskovec 2011]. As shown, AUC values are surprisingly stable even when the sampling parameters are radically changed, while the precision drops after a while if connectivity and activity decrease. This means that the quality of the similarity ranking gets worse, but not enough to decrease the AUC, thus confirming the goodness of the prediction.

5.3.1 *Discussion.* The overall picture that emerges from the experiments reveals some interesting results. First of all, the strong correlation between social linking and user activity, resulting in a noticeable homophily phenomenon, can be profitably exploited to accurately infer the structure of the social network given only information on user features. Considering various features results in quite different prediction performance, as seen in Figure 14.

The only disadvantage of social link prediction based on folksonomic information is that in many systems a considerable portion of users does not use tags: 50% of users are taggers in Last.fm, only 30% in aNobii. However, when tagging information is available, results are encouraging. For folksonomy-based features (items and tags), our prediction methodology based on the MIP similarity metric tends to work better than the other similarity measures considered. This result holds across different collaborative tagging systems. When users are active (described by a high number of triples), we observe that good results are achieved with a distributional approach when aggregating over tags and with a collaborative approach when aggregating over items. In other words, from a collaborative filtering perspective, knowing that two users share a tag is more informative for predicting their social link. In the distributional scenario with item representation, the accuracy compares favorably with that achieved by the Last.fm tasteometer.

It is interesting to notice that the ranking of aggregation methods by prediction accuracy is not consistent across datasets. For example, in the Most Connected scenario, the distributional and the collaborative approaches behave differently in the two datasets considered (see Table III). This means that even folksonomies with the same macro-structural properties (broad folksonomies, with similar numbers of users, triples and tags) can be characterized by inequivalent tagging patterns that lead to different performance of the prediction techniques applied to them.

Predictions made from the group feature can lead to even more accurate results.

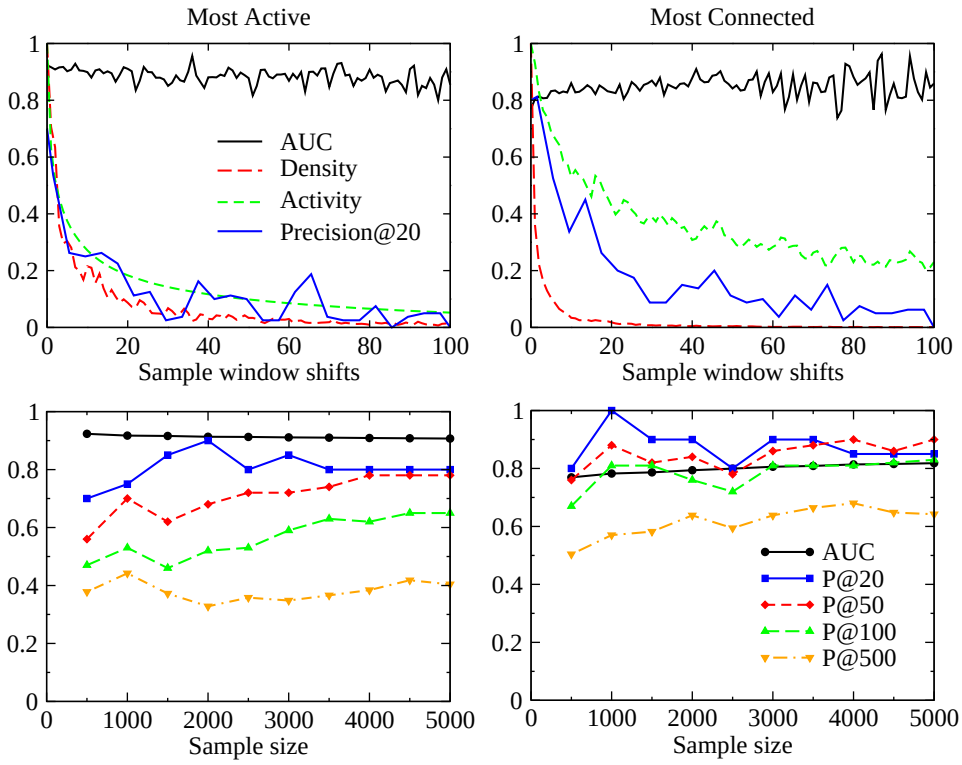

Fig. 15. Sensitivity analysis for the library feature in aNobii; the Most Active (left) and Most Connected (right) cases are considered. Top plots show AUC and precision at 20 as the sample window is shifted. Each step represents a shift of the sample window of 250 positions on the rank of most active/connected users; the window size is left unchanged. The decay of the density of the social graph and of the activity (number of books), normalized on their initial values, are shown jointly. The bottom plots shows how AUC and precision at N changes as the sample window size is increased up to 5,000 users.

Groups behave very well if the total number of groups in the system is not too small with respect to the user population; furthermore, compared to the number of taggers, a larger portion of users in social systems take part in thematic groups, thus allowing the prediction for a wider user set. Even if we do not focus here on the causality aspects linking social connections with homophily, the high AUC values obtained for the group feature could reasonably lead us to conjecture that groups are effective means of socialization, i.e., people know each other through groups.

The best performing profile feature is the library. Aside from the surprisingly high accuracy obtained for the Most Active set, where the information is maximal, the most important outcome is that the analysis of library feature vectors is very significant also in cases when considerably less information is available. From this viewpoint, the Most Connected scenario in Last.fm is particularly revealing because the accuracy is very high even if users are described with feature vectors containing at most 50 artists from their libraries. In a nutshell, the prediction task performs best if it relies on the main feature that denotes the social network topic; in our

Table IV. Predictive power of single and combined features (J48 decision tree) on a balanced set of 10,000 positive and negative samples extracted from the aNobii dataset.

Feature	Tags collab MIP	Groups	Library	CN	Profile feaures	All features
AUC	0.785	0.807	0.811	0.844	0.924	0.963

case study, books for aNobii and artists for Last.fm.

Finally, the common neighbors baseline seems to be a very good predictor of social links, which sometimes performs even better than all the other profile features. This is in part expected because the common neighbors measure captures the user-to-user similarity due to their probability to form a *triadic closure*, which is a relevant phenomenon of attachment in social networks [Newman 2001].

Next, we explore an hybrid approach that tries to properly combine profile features and network-based features to obtain an even greater accuracy.

5.4 Combining features for prediction

The common approach to link prediction based on a multiple feature set relies on machine learning techniques. Prediction is seen as a binary classification problem that can be solved with a classifier, trained on the features that describe the nodes. Positive and negative samples are chosen among pairs of nodes that are connected or disconnected, respectively [Lü and Zhou 2010].

Here we adopt this approach by selecting $10,000$ positive samples and as many negative samples from the set of aNobii users who have at least one instance of each feature in their profile and with at least one outgoing edge. The features considered are simply the similarity scores computed as described previously in this Section. We used the J48 decision tree from WEKA [Hall et al. 2009] as binary classifier and we performed a 10-fold cross validation on our sample. We run the classifier for each profile feature and for the common neighbors separately, then we combine together all the profile features only and finally we add also the information of common neighbors similarity. As a measure of accuracy, we coherently keep considering the AUC.

The results are shown in Table IV. The main observation that should be pointed out in addition to previous discussion is that combining different features results in a noticeable boost of the predictive power. In particular, the combination between different profile features gives about a 35% improvement over the best performing profile feature taken individually and adding some topological information like the number of common neighbors leads to an additional 10% improvement.

In conclusion, even if predicting links from network-based feature can lead to accurate results, this analysis proves that network-based features can be far more predictive if they are combined with profile features. So, we confirm that profile features are crucial in the link prediction task.

6. LANGUAGE COMMUNITY ANALYSIS

The very high accuracy of our social link predictions in aNobii motivated us to further inspect the reasons behind such a strong performance. We suspected that the results were somehow influenced by the strongly clustered structure of the aNobii social network.

In fact, aNobii is split among two main groups: the Italian community (about 60% of users), and the Far East community, representing Hong Kong and Taiwan (about 20% of users). Since the type of literary items consumed by the great majority of people is strictly entangled with their mother tongue, the intersection of topical interests between the two communities is very small and prevalently limited to few worldwide best sellers. In addition to this, aNobii allows the insertion of annotations with any language character set. As a result, users fill their libraries with books written in their own language and they are motivated to annotate them using terms from their mother tongue vocabulary. Given the topical similarity between neighbors (see Section 4), and given that the two communities have very different topical interests, these two main groups turn out to be almost disconnected in the social network.

We show that this scenario directly influences the performance of our feature-based prediction method, considering as a representative example the Most Active sample of users. The same qualitative considerations hold for other samples and features. We split the Most Active set of taggers into clusters on a country level (within the set of top 500 taggers, 249 are Italian and 137 are Taiwanese) and we calculate some basic measures for pairs of users that reside in the same country or in different countries (we neglect the users that do not specify a location in the profile).

For every user pair, we compute the cosine similarity between their vocabularies and between their item sets, and we measure the portion of inter- and intra-cluster links. To show that the portion of links residing inside a language community is not simply due to statistical properties caused by the size imbalance between different communities, we repeated the same measure on a shuffled version of the social graph, where each node keeps its out-degree but rewires its links at random. Statistics are summarized in Table V. We notice that in aNobii the portion of inter-community links in the real network is considerably smaller than in the shuffled network, meaning that the clusters are nearly disconnected from each other due to language homophily. Furthermore, on average, the similarity between pairs of users, calculated for both tags and items, is much lower for users belonging to different geographic clusters compared to pairs of users that reside inside the same cluster.

Table V. Statistics on language communities that compose the Most Active users who declare country of origin in aNobii and Last.fm. We report on the portion of links that reside inside a cluster or, conversely, connect users belonging to different clusters in the real social network vs. its shuffled version (in parenthesis). The average cosine similarity for tag and item sets computed between pairs of users residing in the same or in different communities is reported as well.

	aNobii			Last.fm		
	Links (shuffl.)	Tags σ	Items σ	Links (shuffl.)	Tags σ	Items σ
Intra	85% (38%)	$3.4 \cdot 10^{-2}$	$2.2 \cdot 10^{-2}$	17% (9%)	$1.4 \cdot 10^{-1}$	$1.7 \cdot 10^{-2}$
Inter	15% (62%)	$4.7 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$	83% (91%)	$1.4 \cdot 10^{-1}$	$1.5 \cdot 10^{-2}$

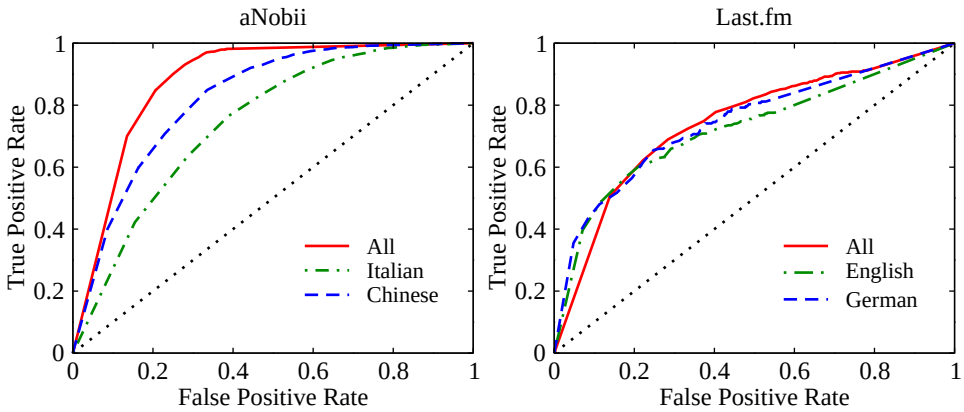

Fig. 16. ROC curves comparing the link prediction within different language communities in aNobii and Last.fm. The user samples considered are composed by the top 500 taggers in the whole system (All) or considering a single language community (Italian, Chinese, English, German). In all cases we used the MIP similarity metric using a distributional aggregation over tags.

The average inter-community topical overlap is thus very small.

The effect of language on facilitating link prediction can be verified by focusing on communities with homogeneous language. We performed the prediction task on subsets of 500 active users within language communities. The comparison between the ROC curves is depicted in Figure 16 for the MIP metric, using the aggregation on tags (the result is qualitatively the same for the other cases). The predictions are significantly more accurate in the mixed community than in the homogeneous communities: the prediction task is simpler in the former case because links between the two communities will almost never be predicted, thus considerably decreasing the number of false positives. Language thus plays a key role in the prediction task for multi-language communities.

To further confirm this observation, we performed the same tests on Last.fm, taking into account the most active users from the two largest language groups in our snapshot: the German community and the English-speaking community, composed by the union of users from USA, UK and Australia. As Figure 16 shows, in this case the prediction accuracy is not clearly affected by the language. Such a different result is well explained by the statistics reported in Table V. Compared to aNobii, the language communities are fuzzier in Last.fm since the level of language homophily is considerably lower and close to that of the random shuffled network; furthermore, there is no clear difference between inter- and intra-cluster feature similarities. Therefore, language does not play a substantial role in the link prediction accuracy.

The influence of language clustering on the tagging behavior in the two social networks considered is also shown in the heat maps in Figure 17. Here, for both aNobii and Last.fm, we computed the cosine similarity between the tag vocabularies in use by the 10 most populated geographic communities. The results clearly show an overall low similarity value between the aNobii communities, except for those countries that share the same language (e.g. Canada, United States and United

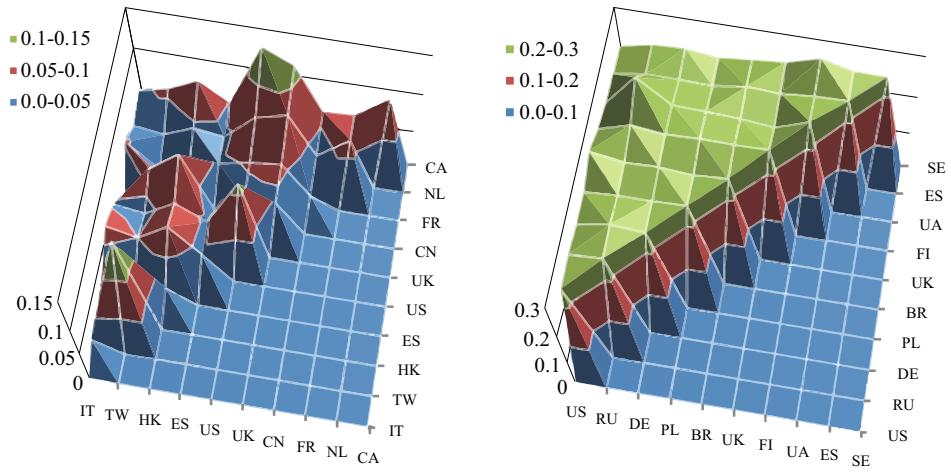

Fig. 17. Cosine similarity between the tag vocabularies used by the 10 biggest geographic communities in aNobii (left) and Last.fm (right). A tag vocabulary is the set of tags obtained by merging all the tag sets of all the users in the community. Color intensity denotes different classes of similarity values.

Kingdom or Taiwan and Hong Kong). On the contrary, the Last.fm geographic groups are more homogeneous, with a substantially higher similarity and no clear distinction on the basis of their official languages. A very similar scenario is found for similarity on item sets (not shown).

Linguistic constraints are more tangible in aNobii than in Last.fm, or in books compared to music. The tagging behavior and the literary tastes of aNobii users are strongly influenced by their language, while in Last.fm, users across different languages tend to have more tags and more music items in common. In synthesis, language can strongly influence the tagging behavior and topical interests of users; the extent to which this happens depends on the topology of the social community and the nature of the objects shared between users. The level of homophily in a social network depends on the extent to which language is correlated with topical interests.

7. CONCLUSIONS AND FUTURE WORK

In this paper we studied the correlation between social and semantic features in three popular online social networks: Flickr, Last.fm, and aNobii. Interesting patterns that are common to all the considered social environments, despite their different magnitudes and semantic orientations, emerge from the study.

We first inspected the interplay between individual user features and social interaction. We observed a strong correlation between the social connectivity and the intensity of explicit user activities like tagging or participation in groups. Assortative mixing patterns between neighbors can be found for all the examined features as well. We showed a dependency between the topical similarity between pair of users and their

shortest-path distance on the social graph, recurring to a null model built to discern purely statistical assortativity from actual homophily. We found a clear topical alignment trend between neighbors for all the examined features.

The results obtained for homophily led us to investigate the possibility of predicting the presence of a social link based only on user profile features. We explored this opportunity in Last.fm and aNobii. We built a user-to-user similarity network for each feature available (tags, tagged items, groups, and library items) using several similarity metrics. We used the similarity rankings of user pairs to perform predictions of social links and we evaluated the predictions' accuracy using ROC curves.

Results shows that all the features have substantial predictive power. Regarding the folksonomy-derived features the similarity metric that we introduced to compare objects in the folksonomy space, namely the MIP similarity, produced the most accurate results. Moreover, the aggregation technique used to extract folksonomic features from the three-dimensional folksonomy space is very relevant to the prediction results. Overall, the library feature, i.e., the collection of user items (books in aNobii and artists in Last.fm) results to be the most predictive, and very accurate even when users have a small number of library items. Finally, combining the feature-based similarity with network-based similarity metric through a machine-learning approach leads to surprisingly good prediction results.

We studied also the influence that confounding aspects like spoken language and nationality have on the prediction task. We showed that the link prediction task is easier in social networks that are strongly clustered by language, because users in different clusters tend to have very different topical interests and to establish very few social ties. For this reason, we suggest that a preliminary analysis on the language community structure should be carried out before testing any feature-based link prediction algorithm on real social network data.

The present findings suggest a number of possible future directions. First, the causal relationships of homophily deserve to be explored in detail. More generally, given longitudinal snapshots of the social environment, evolutionary patterns of the interplay between connectivity and similarity could be studied. Using temporal datasets would also allow better testing of the validity of our link prediction approach by checking if predicted links actually appear in the near future.

ACKNOWLEDGMENTS

We are grateful to Flickr, Last.fm, and aNobii for making their data available. This work has been partly supported by the project *Social Integration of Semantic Annotation Networks for Web Applications* funded by National Science Foundation award `IIS-0811994` and by the Italian Ministry for University and Research (MIUR), within the framework of the project "Information Dynamics in Complex Data Structures" (PRIN). R. Schifanella was supported by the World Wide Style project (WWS) of the University of Turin. Finally we acknowledge support from

the Indiana University School of Informatics and Computing and the ISI Foundation in Torino, Italy.

REFERENCES

Aiello, L. M., Barrat, A., Cattuto, C., Ruffo, G., and Schifanella, R. 2010. Link creation and profile alignment in the anobii social network. *Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust, 2010 IEEE International Conference on 0*, 249–256.

Aral, S., Muchnik, L., and Sundararajan, A. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences 106,* 51, 21544–21549.

Backstrom, L. and Leskovec, J. 2011. Supervised random walks: Predicting and recommending links in social networks. In *WSDM '11: Proceedings of the 4th ACM international conference on Web search and data mining*. ACM.

Benchettara, N., Kanawati, R., and Rouveirol, C. 2010. Supervised machine learning applied to link prediction in bipartite social networks. In *Social Network Analysis and Mining, International Conference on Advances in*. IEEE Computer Society, Los Alamitos, CA, USA, 326–330.

boyd, d. 2009. Streams of content, limited attention: The flow of information through social media.

Caragea, D., Bahirwani, V., Aljandal, W., and H. Hsu, W. 2009. Ontology-based link prediction in the livejournal social network. In *SARA'09: Proceedings of the 8th Symposium on Abstraction, Reformulation and Approximation*.

Catanzaro, M., Boguñá, M., and Pastor-Satorras, R. 2005. Generation of uncorrelated random scale-free networks. *Phys. Rev. E 71*, 027103.

Cattuto, C., Benz, D., Hotho, A., and Stumme, G. 2008. Semantic grounding of tag relatedness in social bookmarking systems. In *Proceedings of the 7th International Semantic Web Conference (ISWC08)*. LNCS, vol. 5318. Springer-Verlag, 615–631.

Clauset, A., Moore, C., and Newman, M. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature 453*, 98–101.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., and Suri, S. 2008. Feedback effects between similarity and social influence in online communities. In *KDD '08: Proceeding of the 14th ACM international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 160–168.

Dunlavy, D. M., G., K. T., and Acar, E. 2010. Temporal link prediction using matrix and tensor factorizations. In *arXiv:1005.4006*.

Fawcett, T. 2006. An introduction to roc analysis. *Pattern Recogn. Lett. 27*, 861–874.

Feldman, R. and Sanger, J. 2006. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Getoor, L. and Diehl, C. P. 2005. Link mining: a survey. *ACM SIGKDD Explorations Newsletter 7,* 2, 3–12.

Getoor, L., Friedman, N., Koller, D., and Taskar, B. 2003. Learning probabilistic models of link structure. *J. Mach. Learn. Res. 3*, 679–707.

Golder, S. and Huberman, B. A. 2006. The structure of collaborative tagging systems. *Journal of Information Science 32,* 2 (April), 198–208.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. 2009. The weka data mining software: An update. *SIGKDD Explorations 11*.

Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M. 2006. Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*.

Haveliwala, T. H. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering 15*, 784–796.

Huan, Z. 2006. Link prediction based on graph topology: The predictive value of the generalized clustering coefficient. In *LinkKDD'06: Proceedings of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Kashima, H. and Abe, N. 2006. A parameterized probabilistic model of network evolution for supervised link prediction. In *ICDM '06: Proceedings of the 6th International Conference on Data Mining*. IEEE Computer Society, Washington, DC, USA, 340–349.

Kumar, R., Novak, J., and Tomkins, A. 2006. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*. ACM Press, New York, NY, USA, 611–617.

Kunegis, J., De Luca, E., and Albayrak, S. 2010. The link prediction problem in bipartite networks. In *Computational Intelligence for Knowledge-Based Systems Design*, E. Hllermeier, R. Kruse, and F. Hoffmann, Eds. Lecture Notes in Computer Science, vol. 6178. Springer Berlin / Heidelberg, 380–389.

Leenders, R. 1997. Longitudinal behavior of network structure and actor attributes: modeling interdependence of contagion and selection. In *Evolution of social networks, Volume 1*, P. Doreian and F. Stokman, Eds.

Lerman, K. and Jones, L. 2007. Social browsing on flickr. In *ICWSM'07: Proceedings of International Conference on Weblogs and Social Media*. http://arxiv.org/abs/cs.HC/0612047.

Leroy, V., Cambazoglu, B. B., and Bonchi, F. 2010. Cold start link prediction. In *SIGKDD'10: Proceedings of the 16th ACM Conference on Knowledge Discovery and Data Mining*. Washington, DC.

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

Leskovec, J. and Horvitz, E. 2008. Planetary-scale views on a large instant-messaging network. In *WWW '08: Proceeding of the 17th International Conference on World Wide Web*. ACM, New York, NY, USA, 915–924.

Li, X., Guo, L., and Zhao, Y. E. 2008. Tag-based social interest discovery. In *WWW'08: Proceedings of the 17th International Conference on World Wide Web*. ACM, New York, NY, USA, 675–684.

Liben-Nowell, D. and Kleinberg, J. 2003. The link prediction problem for social networks. In *CIKM'03: Proceedings of the 12th International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 556–559.

Lin, D. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, J. W. Shavlik, Ed. Morgan Kaufmann, 296–304.

Lü, L. and Zhou, T. 2009. Role of weak ties in link prediction of complex networks. In *CNIKM '09: Proceeding of the 1st ACM international workshop on Complex networks meet information and knowledge management*. ACM, New York, NY, USA, 55–58.

Lü, L. and Zhou, T. 2010. Link prediction in complex networks: A survey. *Preprint*, http://arxiv.org/abs/1010.0725.

Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., and Stumme, G. 2009. Evaluating similarity measures for emergent semantics of social tagging. In *WWW'09: Proceedings of the 18th International Conference on World Wide Web*. ACM, New York, NY, USA.

Markines, B. and Menczer, F. 2009. A scalable, collaborative similarity measure for social annotation systems. In *HT'09: Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*. ACM, New York, NY, USA.

Markines, B., Roinestad, H., and Menczer, F. 2008. Efficient assembly of social semantic networks. In *HT'08: Proceedings of the 19th ACM Conference on Hypertext and Hypermedia*. ACM, New York, NY, USA, 149–156.

Marlow, C., Naaman, M., boyd, d., and Davis, M. 2006. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the 17th Conference on Hypertext and Hypermedia*. ACM, New York, NY, USA, 31–40.

Maslov, S., Sneppen, K., and Zaliznyak, A. 2004. Detection of topological patterns in complex networks: correlation profile of the Internet. *Physica A 333*, 529–540.

McPherson, M., Smith-Lovin, L., and Cook, J. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology 27*, 415–444.

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

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. 2007. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, New York, NY, USA, 29–42.

Mislove, A., Viswanath, B., Gummadi, K. P., and Druschel, P. 2010. You are who you know: inferring user profiles in online social networks. In *WSDM '10: Proceedings of the 3rd ACM international conference on Web search and data mining*. ACM, New York, NY, USA, 251–260.

Molloy, M. and Reed, B. 1995. A critical point for random graphs with a given degree sequence. *Random Struct. Algorithms 6*, 161–179.

Murata, T. and Moriyasu, S. 2007. Link prediction of social networks based on weighted proximity measures. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, Washington, DC, USA, 85–88.

Newman, M. E. J. 2001. Clustering and preferential attachment in growing networks. *Physical Review E 64*, 025102.

Newman, M. E. J. 2002. Assortative mixing in networks. *Physical Review Letter 89*, 208701.

Newman, M. E. J. 2003. Mixing patterns in networks. *Physical Review E 67*, 026126.

Newman, M. E. J. and Park, J. 2003. Why social networks are different from other types of networks. *Phys. Rev. E 68,* 3 (Sep), 036122.

Pastor-Satorras, R., Vázquez, A., and Vespignani, A. 2001. Dynamical and correlation properties of the Internet. *Phys. Rev. Lett. 87*, 258701.

Popescul, A., Popescul, R., and Ungar, L. H. 2003. Structural logistic regression for link analysis. In *Proceedings of the Second International Workshop on MultiRelational Data Mining*.

Prieur, C., Cardon, D., Beuscart, J.-S., Pissard, N., and Pons, P. 2008. The strength of weak cooperation: A case study on flickr. Tech. Rep. arXiv:0802.2317v1, CoRR.

Salton, G. 1989. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Santos-Neto, E., Condon, D., Andrade, N., Iamnitchi, A., and Ripeanu, M. 2009. Individual and social behavior in tagging systems. In *HT'09: Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, C. Cattuto, G. Ruffo, and F. Menczer, Eds. ACM, New York, NY, USA, 183–192.

Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F. 2010. Folks in folksonomies: social link prediction from shared metadata. In *WSDM '10: Proceedings of the 3rd ACM international conference on Web search and data mining*. ACM, New York, NY, USA, 271–280.

Serrano, M. A. and Boguñá, M. 2005. Tuning clustering in random networks with arbitrary degree distributions. *Phys. Rev. E 72*, 036133.

Shalizi, C. and Thomas, A. 2010. Homophily and contagion are generically confounded in observational social network studies. *preprint*, arxiv:1004.4704.

Stäger, M., Lukowicz, P., and Troster, G. 2006. Dealing with class skew in context recognition. In *Proceedings of the 26th IEEE International Conference on Distributed Computing Systems Workshops*. 58.

Szell, M., Lambiotte, R., and Thurner, S. 2010. Multirelational organization of large-scale social networks in an onli ne world. *Proceedings of the National Academy of Sciences 107,* 31, 13636–13641.

Taskar, B., Wong, M. F., Abbeel, P., and Koller, D. 2003. Link prediction in relational data. In *NIPS'03: Neural Information Processing Systems Conference*. Vancouver, Canada.

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

Vázquez, A., Pastor-Satorras, R., and Vespignani, A. 2002. Large-scale topological and dynamical properties of the Internet. *Phys. Rev. E 65*, 066130.