Aiello et al. *EPJ Data Science* 2012, **1**:12
http://www.epjdatascience.com/content/1/1/12

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Link creation and information spreading over social and communication ties in an interest-based online social network

Luca Maria Aiello[1*], Alain Barrat[2,3], Ciro Cattuto[3], Rossano Schifanella[1] and Giancarlo Ruffo[1]

*Correspondence: aiello@di.unito.it
[1]Department of Computer Science, University of Torino, Torino, Italy
Full list of author information is available at the end of the article

Abstract

Complex dynamics of social media emerge from the interaction between the patterns of social connectivity of users and the information exchanged along such social ties. Unveiling the underlying mechanisms that drive the evolution of online social systems requires a deep understanding of the interplay between these two aspects. Based on the case of the aNobii social network, an online service for book readers, we investigate the dynamics of link creation and the social influence phenomenon that may trigger information diffusion in the social graph. By confirming that social partner selection is strongly driven by structural, geographical, and topical proximity, we develop a machine-learning social link recommender for individual users trained on a set of features selected as best predictive out of several and we test it on the still widely unexplored domain of a network of interest. We also analyze the influence process from the two distinct perspectives of users and items. We show that link creation plays an immediate effect on the alignment of user profiles and that the established social ties are a good substrate for social influence. We quantitatively measure influence by tracking the patterns of diffusion of specific pieces of information and comparing them with appropriate null models. We discover an appreciable signal of social influence even though item consumption is a very slow process in this context. All the detected patterns of social attachment and influence are observed to be stronger when considering the social subgraph on which communication effectively occurs. Based on our study of the dynamics of the aNobii social network, we investigate the possibility to predict the evolution of such a complex social system.

1 Introduction

Global dynamics of online social media emerge from the aggregation of the behavioral footprints generated by the activity of the users and their interactions. Such complex information ecosystems are characterized by two fundamental components, namely the creation of *social connections* between individuals and the *information exchange* between them. Mining the static and evolutionary patterns of such phenomena is the key to understand and predict micro and macroscopic dynamics of the whole system.

So far, many efforts have been focused on investigating the causes that determine the creation of social links and the process of information diffusion along these links. If, on the one hand, some results obtained by previous work are supported by well-known soci-

ological theories, on the other hand many dynamics characterizing online social systems are not intuitive, difficult to model accurately, and still widely unknown.

Among others, the microscopic dimension of the process of *link creation*, and the *influence* phenomenon that triggers the diffusion of a piece of information or the spreading of a behavioral norm across the social network have still many unexplored sides. In the first case, even though many studies have addressed the problem of predicting the global evolution of social graphs, only few investigations have been performed from the individual perspective, namely trying to predict future social connections of a single social agent. Similarly, even if several models of information spreading on social networks have been proposed in the past, it is still not clear to what extent information can spread quickly and effectively in the network, and whether the factors that determine influence between peers are generalizable across different social systems.

We contribute to shed light on these questions through the analysis of aNobii, an online social network for book lovers. Unlike the mainstream, general-purpose social networks (*e.g.*, Facebook, Twitter, Google+), aNobii is a *network of interest*, where social aggregation is determined by the topical interests of the readers. Moreover, the contact network can be retrieved *via* crawling without restrictions, thus allowing the analysis of all the nodes reachable *via* crawling. The specificity of the domain considered and the richness of the features publicly exposed by the users allow the exploration of the social dimensions from an unusual angle. Our analysis is driven by two main goals:

1. Designing an effective strategy for the *recommendation* of new links to single users and verifying its effectiveness in an interest-based network. As opposed to the task of link prediction, link recommendation is a widely unexplored task and it has been addressed only for general-purpose networks. We survey a large amount of structural and topical features and we determine the best ones for recommendation purposes. We verify that recommendation in the considered domain is a harder task if compared to general purpose networks and we provide insights on the origin of this difference.
2. Providing quantitative measure of influence in an interest-driven domain by investigating the perspectives of both users (pairs of individuals interacting and exerting influence on one another) and items (books spreading in the network by word-of-mouth process). In the analysis of book spreading, we provide a novel comparison of diffusion traces with null models, we detect a clear signal of influence in a domain in which the consumption of items is a slow process, and we highlight some factors that foster the adoption of books by individuals.

A number of results emerge from the present study, including:

- The analysis of static properties of the aNobii online social system, including geographical and topical bias in link connectivity;
- The discovery of a relation between an item popularity and its viral diffusion potential;
- The detection of the importance of communication patterns over mere social links in the process of social link creation and influence;
- The introduction of a metric of structural node similarity inspired by graph-centrality metrics.

Overall, we present here a comprehensive study of the structure and dynamics of a social system that can be a valuable reference in online social network analysis, as it proceeds all the way from the data collection to the study of the complex dynamics of the system.

In the following we give an overview of the related work in the field (Section 2), we introduce the details of the aNobii social network and we discuss its static and dynamic structural properties (Sections 3, 4). The dynamics of link connectivity and social influence are discussed in Section 5, and the link recommendation algorithm is presented in Section 6.

2 Related work

Several studies have described online social systems from the point of view of their static network properties [1] and in the dynamics of their overall evolution [2, 3]. Temporal fluctuations of network topological features such as diameter, clustering coefficient and mixing patterns [4] and dynamics of link creation in social networks [5] have been explored in depth through the analysis of large-scale real world datasets. Previous works on link characterization, focusing on the patterns that describe the creation of links and how social ties features evolve in time, reveal that link creation is driven by *proximity*, *triangle closure*, *reciprocation* and *homophily* [5–9].

Among the topics related to the analysis of multi-agent systems, in this paper we focus on three broad areas that have attracted a strong interest; namely, the study of the *communication* patterns between users, the *influence* phenomenon, and the *link prediction* problem.

2.1 Communication networks

Recently, findings from social network analysis have been corroborated and expanded by the study of communication networks - also denoted as *activity* networks [10] or *interaction* networks [11] - that often coexist with social networks. The comparison of the graph of user-to-user interactions with the social network reveals similar connectivity patterns driven by reciprocity and triangle closure [12].

Activity networks are more dynamic than social networks and reflect changing trends in user interaction and information flow. Communication graphs have shown to be strongly clustered and to change over many time scales, even if the structural features of the activity network remain stable over time [13]. It has been observed that the average interaction level with neighbors in the social network is very low [14] and often decreasing with time [13]; in agreement with this, studies on the Facebook interaction graph [11] reveal that the social links that are effectively exploited for user-to-user communication are a minority. Moreover, recent studies on Twitter revealed that users can entertain no more than 100-200 stable relationships among all their social contacts [15]. Such results confirm the intuition that online social ties are not always good proxies to extract information exchange patterns.

Although the importance of communication links has been assessed in the past, many social phenomena such as homophily and influence have been studied using the graph of conventional social ties as reference. The effectiveness of modeling and predicting some social phenomena using communication networks instead of social networks has not been explored thoroughly, and has not been considered in the case of networks of interest like aNobii. We compare social and interaction ties in the context of link creation and information diffusion, finding that the information they carry have a different potential in predicting the formation of new links or the diffusion of information. For the link recommendation task, we introduce a metric that combines the information from both social and interaction networks to enhance the prediction accuracy.

2.2 Influence and diffusion

The task of capturing the dynamics of information spreading and influence that occur in networked environments has received much attention recently. Diffusion models of word-of-mouth processes have been developed in the past to enhance viral marketing strategies [16]; more recently, due to the large diffusion of social media, detection of influence patterns and of influential individuals has become important to capture the interaction dynamics in social networks and in real-time information networks.

Analysis of information propagation in Flickr [17] showed that diffusion is limited to individuals who reside in the close neighborhood of the seed user and the spreading process is very slow. Analysis of message cascading on Twitter has been used to estimate the degree of influence of users [18]; the most influential among a pair of users is determined using the difference between some activity metric, like the number of followers or number of tweet replies. In partial disagreement with this study, it has been shown that the number of followers (or of social contacts in general) does not imply a high influence degree [19].

A crucial task in the analysis of influence patterns is to discern real influence from unobserved factors, like homophily or confounding variables, that can induce statistical correlation between the behaviors or the profiles of connected users even without one being influenced by the other. Shuffling or randomization tests on user features are commonly used to detect a signal of influence inside noisy patterns of correlation between pairs of users [20]. Investigations on the interplay between homophily-driven creation of social connections and the influence that neighbors exert on each other's behavior have been made by Crandall *et al.* [21] on the Wikipedia collaboration network. Bakshy *et al.* [22] have reported a large scale experiment performed on the Facebook social network by randomizing the exposition of users to the items published by their friends, in order to expose the role of the social links in the propagation of an information, and to show the existence of a genuine influence phenomenon between Facebook friends.

Instead of representing the influence as an infection phenomenon between connected individuals, Yang and Leskovec [23] recently proposed a linear influence model which is agnostic on the network structure and relies only on the time of the contagion. These observations imply the presence of a hidden contagion web that is different from the observed social network [24]. Based on similar observations, other probabilistic models that represent influence effects between peers disregarding social links structure have been proposed [25].

To complement previous studies, we focus here on the influence phenomenon in a social context where item consumption (reading books) is a much slower process than in general-purpose online social and news media. We explore the influence process both from the point of view of users and items, exposing strong signals of influence at the moment of social link creation and the generation of information fluxes over the existing links. We characterize the spreading traces (*i.e.*, graphs of item adopters expanding in time), compare then to null models and provide some insights into the still open question of whether the fraction or the number of influencing neighbors has a stronger impact on the diffusion probability.

2.3 Link prediction

Predicting the presence of a link between two entities in a network is one of the major challenges in the area of *link mining* [26]. Such edge-related mining task is usually defined as

link detection [27] when it aims to disclose the presence of unobserved or unknown links on a static network or as *link prediction* when it aims to foresee whether a connection will arise in the future between nodes that are unlinked at the current time. *Link recommendation* finally is a task whose goal is to provide to a target user a list of contacts that he will likely be keen to form a social link with [28].

Seminal work on link prediction was presented by Liben-Nowell and Kleinberg [29, 30]. They identify structural properties of the graph which can be used to build a ranking of the node pairs based on their structural similarity, which is in turn exploited to predict future interactions. Several slight variants of this approach have been adopted [31]. Another early work by Popescul *et al.* [32] focused instead on link detection using a classifier trained on the *feature vectors* that describe the nodes of the graph.

Combining structural graph similarity measures and simple node-based features in a supervised learning approach to link prediction has been also tried in the past [9, 33], showing the improvement of the prediction performance compared to predictors based solely on topological features. Geographical proximity between nodes [34] and groups affiliation [35] have been effectively used as node-based feature as well. Recently, some tests have been done also on the predictive power of some network clustering algorithms in link prediction tasks [36].

The best-known topological measures of structural similarity between pairs of nodes are reviewed and refined by Zhou *et al.* [37] and Lü *et al.* [38]. The authors compare several structural similarity metrics for link prediction in terms of accuracy and computational efficiency. Novel local proximity measures are also proposed and shown to be efficient and accurate in link detection. Efficiency of structural proximity metrics on graphs is addressed also by Song *et al.* [39].

Detection of links based only on the information extracted from folksonomies is performed by Schifanella *et al.* [40]. Similarity measures explicitly designed for the folksonomic space are used to compute a lexical proximity between users. A similar context is considered by Leroy *et al.* [41], who leverage the group membership information from Flickr to build a probabilistic graph and detect the hidden social graph with a good accuracy.

The problem of detecting both unknown links and missing node attributes in a network is addressed by Bilgic *et al.* [42]. They propose an iterative method that refines at each step the prediction of one of the two features considered leveraging the information gained on the other feature at the previous step.

The role of temporal aspects in prediction is explored by Tylenda *et al.* [43], who exploit the information of recent interaction between individuals to improve the prediction accuracy. Dunlavy *et al.* [44] use a matrix-tensor method to predict links that will be created in the future in networks with an underlying periodic structure.

Even if the majority of papers is focused on link prediction on simple graphs, a few techniques have been developed also for different kinds of networks. Work has been made in link detection on weighted networks [45–47], bipartite networks [44, 48, 49] and signed social graphs [50]. Very recently, an approach that combines supervised learning and random walks has been shown to have a promising accuracy for both prediction and recommendation of new links [28].

Finally, some approaches based on probabilistic models such as relational Markov networks [51] and probabilistic relational models [52] deserve to be cited. These approaches

have however not been proven to be scalable and they have not been extensively tested on real-world datasets.

Despite the large amount of work in the prediction area, few efforts have been devoted to the task of link recommendation, which is inherently different (and more relevant for real social media services) since it aims to the satisfaction of single users and not just to maximize the ability of predicting the global evolution of the social graph. Moreover, to the best of our knowledge, link recommendation has not been studied in networks of interest but only in general-purpose online social graphs like Facebook. In our recommendation method we collect all the most relevant state-of-the-art features used for link prediction, define an additional feature, and we rank them according to their effectiveness in the recommendation task.

3 aNobii dataset

We analyze a temporal dataset taken from aNobii.com, a website for book lovers. The main feature of aNobii is the personal digital *library* that every user can build by picking titles from a vast database of more than 30 millions publications along with their metadata (such as author, publication year, *etc.*). Every book in the library can be marked with a reading status (*e.g.*, 'finished reading') and can be annotated with keywords (tags), a rating (from 1 to 5 stars) and a review. There is also a *wishlist* containing titles that users have planned to read. Users can enrich their profile with other personal information like their gender, age, marital status and a geo-location composed by a country and, optionally, a town. Country is specified in 97% and city in roughly 40% of the profiles.

Channels of social interaction form another crucial component of aNobii. The social network is composed by two different kinds of mutually-exclusive ties, namely the *friendship* and the *neighborhood* relations. Even if it is up to the users to choose one or another, the aNobii website suggests to establish a friendship tie with people that you already know in real life, while neighborhood should be used for people that you do not know, but whose library you consider interesting. Except for this usage recommendation, the two types of link have the same characteristics. They are directed, they can be established even without the approval of the linked user, and they enable the notification of the linked library updates. Social aggregation can be achieved also through the affiliation to *groups*. Thematic groups can be created by any user and the membership is open to anyone. The last channel of interaction is the *message wall* (also called 'shoutbox'). Users can write messages on the walls of any other individual, independently of the existence of a relationship in the social network. Self-posting is also allowed. Message exchange defines a different social network that we call *communication graph*, and whose properties are discussed in detail in Section 4.3. Self-posting is also allowed, yielding self-loops in the communication graph.

We explored the aNobii social networks through *web crawling* and collected all the public user data through page scraping. We took several snapshots, 15 days apart, using a BFS strategy initialized with a random seed and expanding the user list following the links of the contacts lists and in the shoutboxes. Since social and communications connections are directed we were only able to collect the information of the largest strongly connected component and the out component.[a] However, we collected the full information of both components, thus avoiding the possible biases related to incomplete sampling of a connected component.

4 Structure and dynamics of social network

4.1 Overview on network structure

Friendship and neighborhood networks have similar global properties, with however some structural differences. As shown in Table 1, both networks have a high percentage of reciprocated links and a strongly connected kernel that includes the majority of nodes. However, the neighborhood network is slightly smaller, denser, and has higher degree centralization [54]. Its size is smaller because neighborhood ties tend not to be used by less active members and it is more centralized because of very popular libraries with many 'followers': the range of variation of the in- and out-degree are broader for neighborhood than for friendship (for the in-degree, the maximal values are 1,708 for neighborhood and 453 for friendship; for the out-degree, the maximal values are respectively 6,537 and 705). These differences reveal that the two social ties are used slightly differently by users, and are in agreement with the intuition that friendship links correspond to individuals the user really knows, while neighborhood links can be established towards any other user whose library seems of interest. In the context of properties that apply in a comparable quantitative and qualitative way to both networks, it is however more convenient to consider the *union* between them. In the following, for simplicity, we will refer to the union network as the aNobii social network.

As a direct result of their structural differences, the diameters of the two networks (computed as the maximum shortest path length) are appreciably different. Still, they are both very high if considered that similar diameter values have been found for many other online social networks with much greater size [1]. The strong geographical clustering of the social network is the main reason behind this feature. The country-level graph of the social network depicted in Figure 1 reveals that the network has two main geographic communities, namely Italy (with roughly 60% of users) and Far East (composed by Taiwan, Hong Kong and China, that include less than 30% of users altogether). Since these two clusters are loosely connected to each other, the network has a dual core structure where connection between the two cores is mostly mediated by smaller communities (*e.g.*, the USA cluster). Paths between individuals residing in different cores are thus longer if compared to a more ordinary single core configuration and, consequently, the diameter is higher.

Table 1 Statistics concerning the friendship and neighborhood networks, their union (*i.e.*, the full social network) and the communication network in April 2011

	Friendship	Neighborhood	Union	Communication
Nodes	126,858	77,356	140,686	80,303
Links	557,258	633,635	1,187,650	574,285
Loops	0	0	0	22,579
Reciprocation	0.60	0.43	0.54	0.61
$\langle k_{\text{out}} \rangle$	4.4	8.2	8.4	7.2
$\langle w \rangle$	-	-	-	1.8
$\langle m \rangle$	-	-	-	12.9
WCC size	121,143	76,760	140,686	75,965
SCC size	81,292	41,063	100,492	38,336
Density	$3.4 \cdot 10^{-5}$	$1.1 \cdot 10^{-4}$	$6.0 \cdot 10^{-5}$	$8.9 \cdot 10^{-5}$
Average SPL	7.3	4.7	5.3	4.8
Diameter	25	15	20	17
Degree centr. [53]	0.0072	0.0875	0.0486	0.0650

SPL = shortest path length; WCC = weakly connected component; SCC = strongly connected component; $\langle k_{\text{out}} \rangle$ average out degree, $\langle w \rangle$ average edge weight (only for the communication network, see Section 4.3), $\langle m \rangle$ average number of messages in the shoutbox. Degree centralization is given by Freeman's formula $C_D = \frac{\sum_{i \in G} (k_{\max} - k_i)}{(|G-1| \cdot |G-2|)}$.

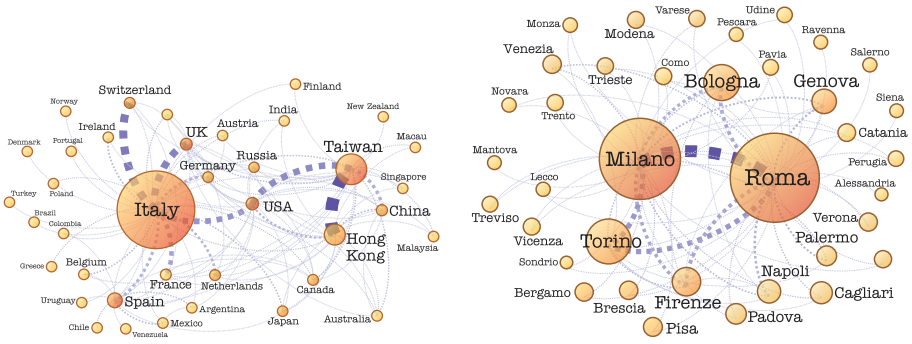

Figure 1 Graphs of home countries and towns of aNobii users. Nodes are scaled according to the size of the communities and the width and colors of edges depend on the number of links that connect nodes between the communities. In the graph of countries, communities with less than 20 members and edges with weight less than 20 are not shown. In the Italian towns graph, communities with less than 100 members and edges with weight less than 100 are not displayed.

The separation between the geographical regions in the graph can be quantified by measuring the *conductance* φ of the graph cut separating the users who reside in a given region R from the rest of the network, and comparing the value with the conductance of a random cut φ^{rand} between a region R' and the rest of the graph, where R' has the same size and degree distribution than R. The conductance is defined as the ratio between the number of edges crossing the cut and the minimum number of edges inside one of the two regions separated by the cut: small values denote well-separated regions while values close to 1 denote strong connectivity between regions [55]. Italy and Far East regions have a much smaller conductance than their random counterparts ($\varphi_{\text{IT}} = 0.08$, $\varphi_{\text{IT}}^{\text{rand}} = 0.69$, $\varphi_{\text{FE}} = 0.05$, $\varphi_{\text{FE}}^{\text{rand}} = 0.24$), while 'bridge' regions have a conductance comparable to the random case ($\varphi_{\text{USA}} = 0.66$, $\varphi_{\text{USA}}^{\text{rand}} = 0.60$).

Narrowing down the view on town-level graphs inside clusters, the intra-cluster connections appear denser and structured around a single core of nodes (Figure 1). Of course, since aNobii is focused on books, language is the main reason that leads to this sharp separation.

In addition to the geographical location, aNobii profiles contain a rich information about users. User activity, along with social ties, can be measured by several indicators. The corresponding probability distributions are shown in Figure 2. Not surprisingly, the most popular activity is filling the library with books (94% of users have at least one book). Approximatively 50% of users added at least one book in the wishlist and roughly the same portion of users is member of at least one group. Books are annotated with reviews by around 40% of users and rated by 75%. Tagging activity is quite unfrequent. Around 75% of users declare at least one friend or neighbor. Moreover, and as expected from studies in other online social systems [56], the activity distributions are all very broad, displaying heavy tails that highlight a strong heterogeneity in the behavior of users: no typical value of any user activity can thus be defined.

Different activities exhibit strong correlations between each other, as also investigated for other social networks like Last.fm and Flickr [40, 56]. Graphically, correlations can be depicted by showing the average activity of users who exhibit a given engagement level for another activity; in Figure 3 we display some correlation graphs of the out-degree and the number of books with other activities. Even if the observed patterns are noisy for users

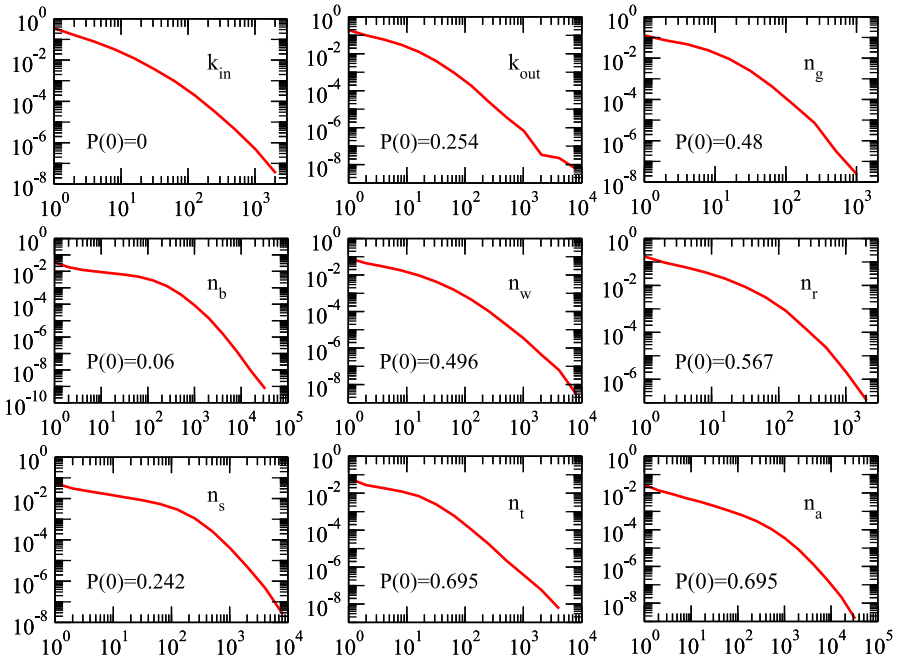

Figure 2 Distributions of the measures of activity of aNobii users. In-degree k_{in} and out-degree k_{out} in the social network, number of group memberships n_g, number of books in library and wishlist (n_b and n_w), number of reviews n_r and books rated with stars n_s, number of distinct tags n_t and total number of tag annotations n_a.

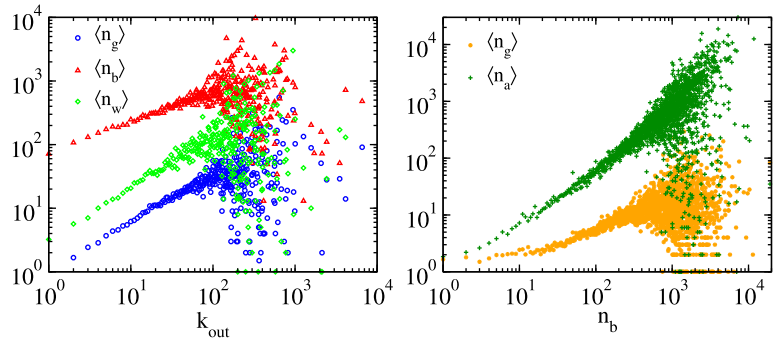

Figure 3 Correlations between different activities in aNobii. Left: Average number of group memberships $\langle n_g \rangle$, of books in library $\langle n_b \rangle$ and wishlist size $\langle n_w \rangle$ against the number of social out-links k_{out}. Right: Average number of group memberships and tag annotations $\langle n_a \rangle$ against the number of books in the library.

with a large number of connections and books (due to the low number of users over which the averages are performed), all the activities considered show a clear increasing trend for increasing values of k_{out} and n_b, corresponding to a positive correlation between activities.

The correlations between the activity of social network neighbors, commonly known as *mixing patterns* [57], can be measured by plotting the average amount of activity of the neighbours of all the users with the same value for that activity (*e.g.*, with the same number of books). The positive slopes of the scatterplots in Figure 4 (modulo the noise for the less frequent activity values) reveal an assortative mixing for all the activities, meaning that

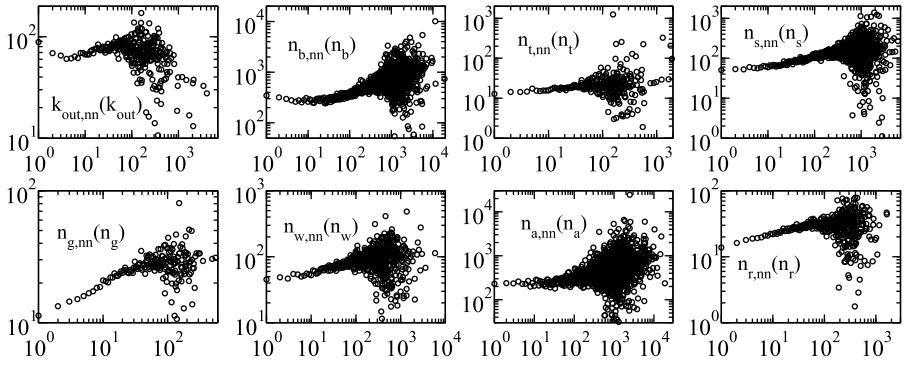

Figure 4 Mixing patterns for different activities in aNobii. The notation $n_{x,nn}(n_x)$ indicates the average amount of an activity x performed by the nearest neighbors of those users that expose the same amount n_x of that activity. As defined before, $k_{\mathbf{out}}$, n_b, n_w, n_g, n_t, n_a, n_r, n_s denote respectively the out-degree and the number of books, wishlist items, group memberships, tags, tag annotations, reviews and ratings.

Table 2 Evolution of some quantities from one snapshot to the next

	1 → 2	**2 → 3**	**3 → 4**	**4 → 5**	**5 → 6**
New nodes	2,241	2,121	1,911	3,214	3,567
Removed nodes	239	222	230	220	684
New edges	19,472	18,324	17,618	24,805	26,883
Removed edges	642	763	713	782	700
New edges existing nodes	10,044	9,296	9,758	11,925	12,520
$u \to v$	54%	53%	54%	55%	51%
Reciprocated	10%	13%	13%	13%	14%
$u \leftrightarrow v$	36%	34%	33%	32%	35%
Simple closure	21%	21%	22%	21%	19%
Double closure	9%	10%	9%	9%	9%

The last two sections report the fractions of different edge types among the new edges created between nodes already existing at the beginning of the time window considered.

users are likely to be linked to other individuals with comparable amount of activity, a typical pattern of social networks.

4.2 Evolution of the network

Our temporal dataset allows to study the evolution of the social network. In Table 2 we report the evolution of some network parameters in a time span of 2 and a half months, with a granularity of 15 days. The largest component grows steadily due to the arrival of new nodes, and new ties are also created between existent users. Node and edge deletion are much rarer events.

We classify newly created edges among existing nodes in three categories. $u \to v$ denotes the category of unidirectional links while $u \leftrightarrow v$ represents the new reciprocal links. 'Reciprocated' denotes instead the new links from a node u to a node v, such that a link from v to u already existed. Links of the type $u \to v$ and $u \leftrightarrow v$ can be further described as 'Simple closure' and 'Double closure' ties respectively if they close at least a directed triangle (*i.e.*, there existed a node w such that the arcs $u \to w$ and $w \leftrightarrow v$ already existed). This fits the expectation that in a social network links are often established toward 'friends of friends'. This *triangle closure* phenomenon is evident also by looking at the distribution at time t of the distances of nodes that become linked at time $t + 1$. The comparison be-

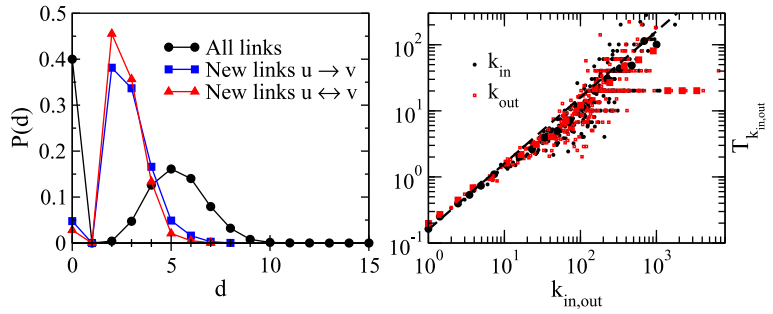

Figure 5 Link creation in aNobii. Left: Distribution at snapshot 4 of the distances of nodes which become linked between snapshots 4 and 5, compared with the distribution at snapshot 4 of distances between all pairs of users. The points in $d = 0$ give the portion of pairs of nodes between which no directed path exists in snapshot 4. Right: Measure of the preferential attachment. T_k is the probability for a new node to create a link toward a node of degree k. Black circles and red squares account for in and out degree, respectively. Thick dots represent the log-binning of the values, and the dashed line represents a linear relationship $T_k \propto k$.

tween such distribution and the distribution of distances between all the node pairs in the network (Figure 5, left) reveals that the process of social partner selection is biased towards the topological vicinity of the user. In particular, more than 40% of the new arcs close triangles and more than 80% are established between nodes residing at distance at most 3.

Besides *triangle closure*, another phenomenon that underlies link creation in growing graphs is *preferential attachment*, *i.e.* users with large number of connection are preferentially chosen to establish a social link [58]. We test this hypothesis using the following method [59]. Let us denote by T_k the *a priori* probability for a newcomer to create a link toward a node of degree k, between time $t-1$ and t. Given that at time $t-1$ the degree distribution of the $N(t-1)$ nodes is $P(k, t-1)$ (*i.e.*, there are $N(t-1)P(k, t-1)$ nodes of degree k), the probability to observe a new link from a new node to a node of degree k between $t-1$ and t is $T_k P(k, t-1)$. Therefore, we can measure T_k by counting for each k the fraction of links created by new nodes that reach nodes of degree k, and dividing by $P(k, t-1)$. As shown in Figure 5 (right), we obtain a linear behavior $T_k \propto k$, both when considering for k the in and the out-degree (which are strongly correlated). This is a clear signal of a linear preferential attachment. T_k values for $k > 1{,}000$ falling far from the diagonal are just statistical noise due to the low number of high-degree nodes.

Clearly, users do not have any knowledge of the overall network topology at any time, so they cannot be more motivated to connect to the most connected users. It is more likely that this preferential attachment arises from the fact that a new user creates links not only towards another user but also towards some of this user's neighbors. It has been shown that this locally-driven connection pattern results in effective preferential attachment [60, 61]. Indeed, we verified in our dataset that many newcomers join the network by creating links to pairs of already connected users.

4.3 Communication and interaction networks

Ties in social media are most often not categorized based on the intensity or on the type of the connections. However, in a social context, ties might have different strength and meaning, depending on the information that flows on them and from the features that describe the individuals they connect. To reach a deeper understanding of social dynamics,

the information on the social connections must be complemented with other relational data. In this respect, the communication network carries a useful information to augment the description of the social substrate as given by the user-declared 'friendship' or 'neighborhood' ties: some user-declared ties might not be the support of any communication, and communication may occur between users that are neither 'friends' nor 'neighbors'.

The most extensive way in which the communication history between individuals can be defined is through a temporal graph, where each edge corresponds to a single message and carries a timestamp. In this temporal graph, the frequency of messages exchanged by two users might change, with periods of inactivity followed by bursts of messages. The detailed study of this dynamics goes beyond the scope of the present study, so that we consider an aggregation over the whole data set time window, and define the *communication graph* as a directed graph where each edge between two nodes is weighted by the number of messages sent between these nodes.

Similarly to previous work [11], we observe that macroscopic structural features of communication graph are analogous to those of the social networks. Degree distributions are very close to those found for the social networks (not shown) and the strength distributions (*i.e.*, number of received or sent messages) reveal an expected broad behavior (not shown). The statistics shown in Table 1 indicate that this graph has high reciprocation and centralization. Note that the communication network has self-links since it is possible for a user to write messages on his/her own shoutbox. Users keeping alive conversation threads on a single shoutbox or announcements published by the shoutbox owner are the main causes of this phenomenon. This behavior concerns however only 28% of the users, and the self-links represent only 4% of the total number of links. In Figure 6 we observe that the social connectivity and the amount of books in the library are correlated with the activity on the communication network. A strong correlation is also found between in-degree and out-degree in the communication network.

As shown in Table 3, the difference between social and communication graphs is substantial. More than 75% of the socially connected pairs lack any form of public communication and, conversely, around 25% of the communication channels are established between non connected users. We call *interaction graph* the portion of the social graph that overlaps with the communication network (*i.e.* Social ∩ Comm in the notation of Table 3).

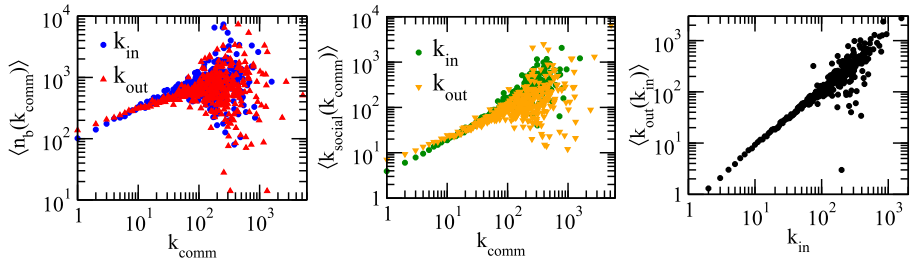

Figure 6 Correlation plots for the communication activity. The numbers of incoming and outgoing messages are strongly correlated with the number of social acquaintances (k_{social}) and with the size of the library (n_b). The high reciprocation in communication implies also a strong correlation between in- and out-degree in the communication network.

Table 3 Overlap between social networks and communication network

		Social\Comm	Comm\Social	Social ∩ Comm
#Nodes	Friendship	57,456	10,901	69,402
	Neighborhood	20,792	23,739	56,564
	Union	63,719	3,336	76,967
#Edges	Friendship	461,774	478,801	95,484
	Neighborhood	435,396	376,046	198,239
	Union	894,946	281,581	292,704

4.4 Topical alignment

Assortative mixing patterns suggest a propensity to the local alignment of behavioral patterns between connected nodes. While we explored only the mixing patterns relative to the amount of activity of neighboring users in Section 4.1, this tendency can be explored more in depth by taking into account the user profiles. More precisely we consider the similarity of users' profiles with respect to specific features, and measuring how it depends on the distance between nodes on the social network. We call *topical local alignment* a static property of the social network for which pairs of individuals that are close in the social graph are more similar than pairs separated by larger distances on the network. For instance, when considering books as a feature, the topical alignment can be measured by computing the similarity between the book sets of pairs of users as a function of their distance on the network. Similarity between two users u and v can be measured by counting the number of common books $n_{cb}(u, v)$ or by considering a normalized similarity measure such as the *cosine similarity*

$$\sigma_b(u, v) = \frac{\sum_b \delta_u(b)\delta_v(b)}{\sqrt{n_b(u)n_b(v)}}, \tag{1}$$

where the indicator function $\delta_x(b)$ is equal to 1 if user x has the book b in his/her library and to 0 otherwise. The cosine similarity is thus a scalar product of the 'book vectors' of users u and v, normalized by the library sizes $n_b(x) = \sum_b \delta_x(b)$.

To check if profiles of neighbors in the social network are topically aligned with respect to some features, we measure the average books and groups similarity of pairs separated by d hops in the social graph; results are shown in the first two columns of Figure 7. A quick decay of the similarity with the distance (for both the cosine similarity and the number of common items) gives a strong clue of the presence of a local topical alignment. However, the detected signal could *a priori* be ascribed to purely statistical alignment effects due to assortativity. For instance, since very active users tend to connect with other highly active users, their similarity could be high just because their feature sets are big, and thus they have a higher chance to share many elements. To tell apart real alignment from statistical effects, we need to compare the results obtained on the real data with a suitable *null model* [40]. Our null model is based on a random reshuffling of the items (*e.g.*, books) between user profiles, keeping unchanged both the size of the item sets and the social connections of each profile. This procedure preserves the assortativity patterns relative to activity intensity but wipes out the alignment due to the interaction between individuals. We note that the randomized curves exhibit a similar decay, that is due to the assortativity effect mentioned above. However, the difference between the curves for the reshuffled and the real data shows that the assortativity alone can not wholly account for the similarity detected in the real data, and that a genuine topical alignment effect is present: the presence

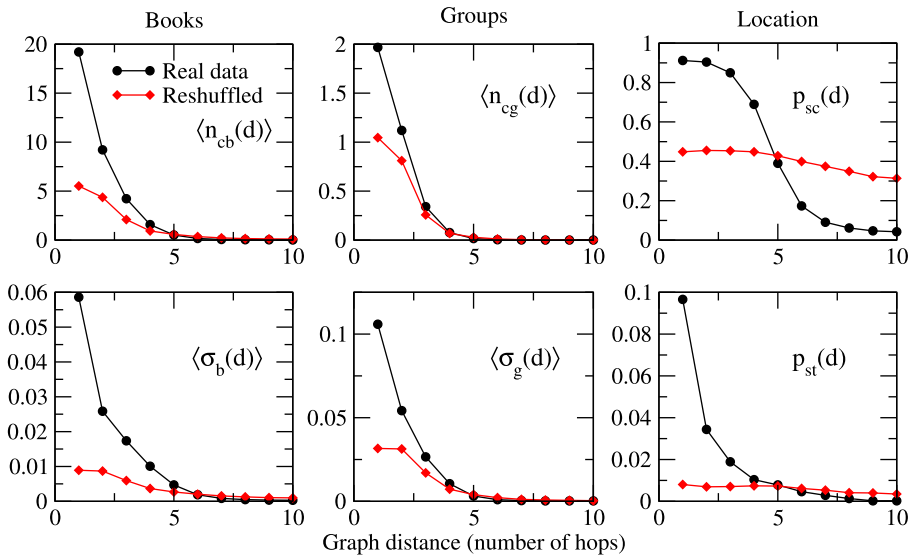

Figure 7 Topical and geographical alignment. Left and middle plots: Average similarity of the libraries (resp., groups) of aNobii users as a function of their distance in the social network. The similarity is measured by the average number of common books or groups (top, $\langle n_{cb} \rangle$, $\langle n_{cg} \rangle$), and by the average cosine similarity (bottom, $\langle \sigma_b \rangle$, $\langle \sigma_g \rangle$) between the books lists (resp., groups). In both cases, the same similarity after random reshuffling of items is shown. Right plots: Fraction of pairs of users at distance d in the union network residing in the same country (P_{sc}) or town (P_{st}). In both cases data from the network with reshuffled links are shown.

of a social link is correlated with the fact that the connected users are more likely to share interests and experiences or to be exposed to the same context or to each other's activity.

The same analysis can be performed on all the features of the users' profiles. For instance, the relationship between the geographic attributes and the distance on the social graph are explored in the right plots of Figure 7 that show the probability that two users at distance d on the social graph are from the same country or town. Again, to disentangle this signal from statistical effects (given for example by the imbalance of the number of users in each nation) we use as null model a random network with the same degree sequence as the original network but reshuffled geographic attributes. The alignment on the nationality feature is strong up to a distance of 4 hops and a strong effect is observed as well for towns, most of all for directly connected users.

This result suggests that people preferentially establish social ties with others who speak the same language, but also that the social selection process is driven by the geographic proximity (*e.g.*, people that reside in the same town). In particular, 90% of the social edges connect users from the same country and there is a 10% probability that two connected users are from the same city. This result indicates a decreasing trend of the probability of connection with geographic distance, as also found in other online social networks that are not based on a particular interest (here, the books) but have broader scopes [62, 63].

As seen in the previous subsection, the fact that two users are connected does not automatically mean that they exchange information through messages. It is therefore of interest to compare the topical alignment on the social links that effectively are the support of communication ('Social ∩ Comm', in the notation of Table 3) with the alignment along the subset of social links on which no communication is observed ('Social\Comm', in the notation of Table 3). Figure 8 shows that the former is larger than the latter, but only slightly:

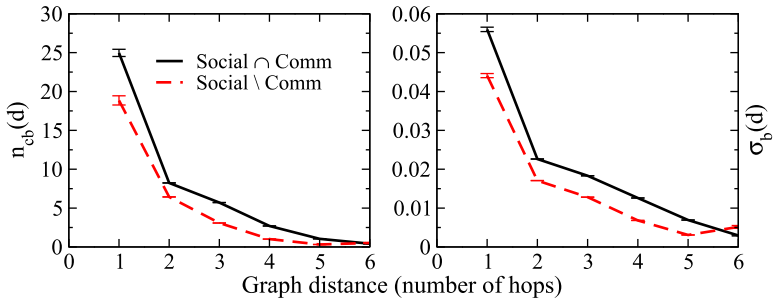

Figure 8 Topical alignment in the interaction graph. Average number of common books (left) and cosine similarity (right) against distance on the interaction graph and on the graph made by social contacts on which no communication occurs (**Social\Comm**). Standard error bars are shown for each curve.

interestingly, strong alignment effects exist even on a network along which no explicit communication flows, and are almost as strong as in the network of communication.

5 Homophily, selection and influence

5.1 Causal connection between similarity and link creation

In Section 4.4 we observed topical alignment as a static property of the network. Here we investigate the evolution of this phenomenon. Since we verified that the topical alignment, which denotes a homophily phenomenon between users [64], is not purely due to assortative patterns, we can ascribe this phenomenon to selection or to social influence. *Selection* corresponds to a process in which the choice of a social partner (here as 'friend' or 'neighbor') that is driven by the similarity between connecting individuals, while *social influence* [20] denotes the tendency of individuals to be influenced in their behavior by others, and in particular by their neighborhood in the social network. As we now show, both phenomena can be exposed in aNobii.

To check whether the occurrence of selection-driven attachment, one needs to compare the topical similarity, computed at time t, between pairs of users who create a social connection between t and $t+1$, with the similarity of another set of users. Choosing for comparison a random set of pairs of users would trivially yield a strong difference, as we have shown previously that (i) most pairs of users creating a link between t and $t+1$ reside at distance 2 or 3 on the network at t, and (ii) the similarity of users at distance 2 or 3 is much stronger than the one of users lying farther apart. We therefore compare the average similarity of connecting users with the average similarity of all the pairs of nodes residing two hops away in the social graph at t. Table 4 shows that pairs of users that are about to get connected are on average more similar than the average over all the nodes that are two hops away (except for the case of the number of common groups averaged over all pairs of users who establish a non-reciprocal link between t and $t+1$). For new bidirectional links, and pairs of users who were at distance 2 and create a link (thus closing a triangle), the average similarity before the creation of the link is particularly strong. This result applies for all the similarity measures considered. The probability that two users at distance 2 have 0 similarity is also much smaller for the users who become linked between t and $t+1$.

The picture emerging from this analysis and from the results presented in Sections 4.2 and 4.4 is the following: users connect to others residing close in the social graph, very

Table 4 Average similarity for snapshot $t = 4$ of pairs forming new links between t and $t + 1$ (either non-reciprocal, $u \rightarrow v$ or reciprocal, $u \leftrightarrow v$), compared with the average similarity of all pairs at distance 2 at t

	$\langle n_{cb} \rangle$	σ_b	$\langle n_{cg} \rangle$	σ_g
$d_{uv} = 2$	9.5 (0.2)	0.02	1.12 (0.61)	0.05
$u \rightarrow v$	12.9 (0.16)	0.04	1.1 (0.6)	0.08
$u \leftrightarrow v$	18.5 (0.06)	0.04	1.67 (0.44)	0.11
Simple closure	18.2 (0.09)	0.04	1.81 (0.45)	0.1
Double closure	23.4 (0.03)	0.05	2.2 (0.36)	0.12

Single and double closure refers respectively to new links $u \rightarrow v$ and $u \leftrightarrow v$ that close triangles. The similarity is measured by the number of common books n_{cb} or groups n_{cg}, and by the corresponding cosine similarities σ_b and σ_g. The numbers in parenthesis give the probability to have similarity equal to 0.

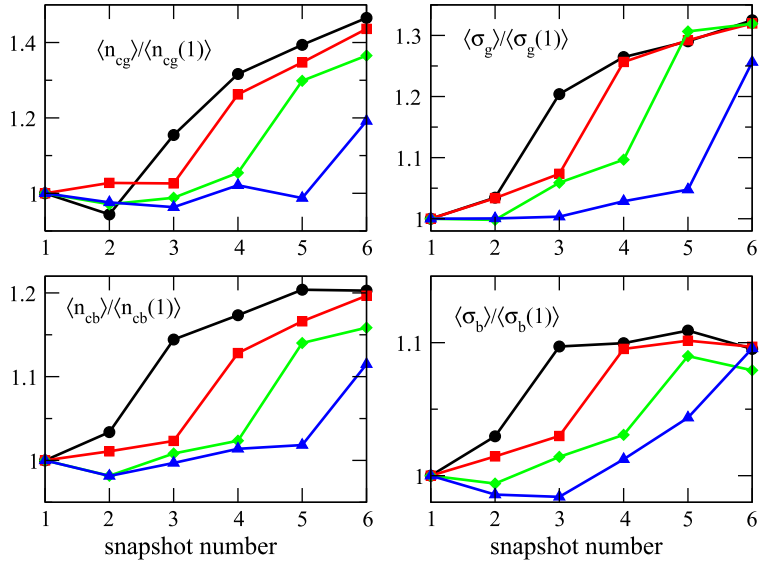

Figure 9 Evolution of the average similarity of user profiles. Similarity, as measured by the numbers of common books or groups, and by the cosine similarity, is shown for links created between t and $t + 1$, for $t = 2$ (black circles), 3 (red squares), 4 (green diamonds), 5 (blue triangles), normalized by the average similarity in the first snapshot. Values are quite stationary before t_0, and clear jumps are observed between t and $t + 1$.

often neighbors of neighbors; moreover, these individuals have on average more similar profiles than other pairs of users at distance 2. In this respect, one can infer that a selection process is at work and is one of the reasons of the observed local topical alignment: among the users who are already close in the graph (distance 2 and 3), the ones who become even closer are the ones who were more similar to each other.

In order to investigate social influence, we instead study the time evolution of the similarity between connected user profiles. In Figure 9 we plot the average similarity for library and group membership features for pairs of users that become connected between t and $t+1$. Before the link is created the similarity score is rather stationary and a sudden increase is observed when the connection is created; the similarity then continues to grow, albeit at a slower rate. The following scenario emerges from this result: after a link creation, newly connected individuals take inspiration from each other for new books to read and new groups to join. The direct consequence of this reciprocal influence is a further alignment of the profiles, *i.e.*, a reinforcement of the homophily.

Note that the similarity metric used is symmetrical, therefore it does not account for the directionality of the newly created link. We decided not to consider the link directionality in the computation of the similarity, as close to 50% of the newly created links are bidirectional (see Table 2), and because a user receiving a new incoming connection is notified about it on his/her personal homepage: the influence at the time of the connection can potentially flow in both directions.

To summarize, our analysis on the dynamics of social aggregations show the presence of a *bidirectional* causal relationship between social connections and similarity. A higher similarity leads to a higher connection probability and, on the other hand, users who get connected become more similar due to the influence that new acquaintances exert on one another. These results apply not only for collaboration networks [21], but also for the present case of interest-based networks such as aNobii, where the similarity between users is evaluated on the basis of profile items, shared metadata, and topics of interest.

5.2 Structure of book graphs

Influence can also be investigated from a different angle, focusing on items rather than on users. The influence observed at the time of a link creation might indeed remain effective for the whole life span of the social link, and, at any time, may lead a user to adopt a new item (in particular a book) from his/her neighbors and, in turn, to influence others to adopt the same item. This phenomenon gives origin to adoption *cascades* that can be studied within the more general scope of information spreading [65]. Better understanding the spreading of items on the network can shed a clearer light on the overall role of influence in the online social network.

In this perspective, we study the static and dynamic properties of the book graphs: a book graph $G(b)$ is defined as the social subgraph composed by the users having the book b in their library or wishlist and by the links between them. We differentiate the analysis by classes of book popularity as measured by the size of the set $A(b)$ of the users who adopted the book b (*i.e.*, the nodes in $G(b)$). In particular, we introduce three popularity classes, namely the *rare* ($|A(b)| \in [10, 500)$), the *middle* ($|A(b)| \in [500, 1{,}000)$), and the *popular* ($|A(b)| \geq 1{,}000$) books. The boundaries of the popularity classes are chosen based on the empirical observation of the popularity distribution of books. Even neglecting very rare books with less than 10 readers we have more than 200K book graphs.

The size of the book graphs are broadly distributed between 1 and 20,000. Book graphs can be formed by several disconnected components, and around 12% of them are composed just by isolated nodes (this is only observed for graphs with at most 100 nodes). The connectivity patterns of books graphs can be detected by measuring how their topological features depend on their size. In Figure 10 we report the relative size of the greatest connected component ($S_{gcc}/|A(b)|$), the relative number of connected components ($N_{cc}/|A(b)|$), and the clustering coefficient (C) against the size of the book graph ($|A(b)|$). For the sake of comparison, for every point in the scatterplot we also depict two twin points representing the same topological measure calculated for two different random graphs taken as null models.

The first is an Erdős-Rényi graph with the same number of nodes and edges. The latter is a random subgraph of the social network with the same number of nodes and the same degree sequence. The purpose of the random subgraph is to model a process in which the book is adopted by the different users at random and independently. If such a

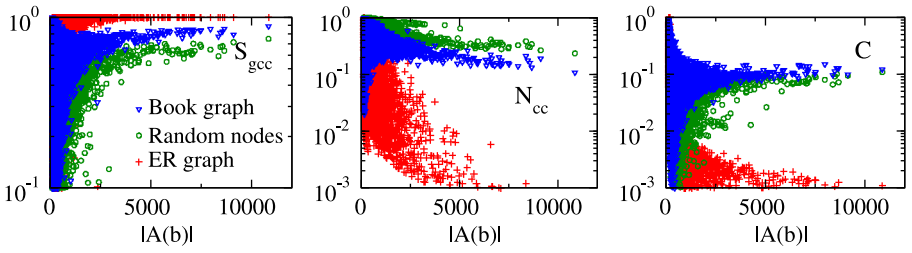

Figure 10 Properties of the book graphs. Scatterplots of relative size of the greatest connected component ($S_{gcc}/|A(b)|$), relative number of connected components ($N_{cc}/|A(b)|$), and clustering coefficient (C) *vs.* number of nodes in the $G(b)$ graph. For each graph, values for an Erdős-Rényi graph with the same number of nodes and edges are reported, as well as for a subgraph of $|A(b)|$ randomly chosen nodes in the aNobii social graph.

process is considered by simply selecting nodes at random in the network, the resulting subgraphs will be almost always composed of isolated nodes or small disconnected components, therefore we impose that the resulting subgraph has the same degree sequence as the original subgraph.

Book graphs exhibit a weaker connectivity but a much more clustered shape than the corresponding ER graphs, at fixed size. The relative number of connected components slowly decreases with size but remains considerably higher than the ER corresponding values; as a consequence, the relative size of the greatest component asymptotically stabilizes around a value smaller than in the ER graphs; conversely, real book graphs are much more clustered. Structural properties of the random-node-graphs are closer to those of the real book graphs, meaning that the measured levels of clustering and connectivity of the book graphs can partly be ascribed to the degree distribution of their nodes. Nevertheless, the random-node model still exhibits lower clustering and higher sparsity than the empirical book graphs.

To investigate more in depth the differences between the random-node model and the real data, we measure the same structural properties at given average node connectivity, and we study the three book popularity classes separately. Figure 11 shows the values of S_{gcc} and C against the average out-degree $\langle k \rangle$ in the book graphs.

The case of ER graphs is the simplest to interpret and is used as a reference for the other two cases. For ER graphs we observe a relatively rapid transition from 0 to 1 for S_{gcc} as $\langle k \rangle$ crosses 1, which is expected given the known transition between a set of small disconnected components and a giant connected component as the probability of connection crosses $1/N$. Instead, the clustering values remain very small (as also expected in ER graphs). In the case of the real book graphs, the size of the greatest component grows smoothly with the average degree, showing no sign of any abrupt transition, suggesting that the connectivity in book graphs is not driven by any threshold mechanism driven by the average node connectivity. Furthermore, for any average connectivity, a non-negligible portion of nodes remains in small isolated components. This can be due to the fact that several users adopt a book independently, without being directly influenced by their online social contacts. However, the clustering coefficient is very large, suggesting that the groups of adopters are tightly knit communities. The random-node model follows the same trend as the real data, but both the size of the largest component and the clustering are lower, showing that the connectivity patterns are not completely due to the degree distribution.

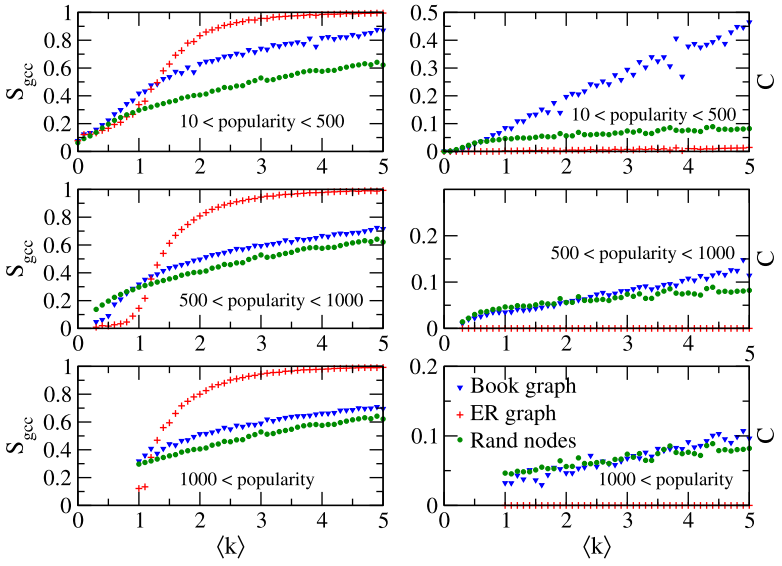

Figure 11 Properties of the book graphs. Relative size of greatest connected components ($S_{gcc}/|A(b)|$) and clustering coefficient (C) *vs.* average out degree ($\langle k \rangle$), averaged at fixed values of $\langle k \rangle$. Comparison values from corresponding ER graphs is resported.

For books with large popularity, the empirical data and the random-node case become closer.

The overall picture tends to indicate that book graphs may be originated by a process of expansion and densification of clustered cores of readers, and that a process of 'contagion' between users might have taken place in the shaping of the subgraphs of adopters $G(b)$. Nevertheless, as the book popularity grows such effect fades, presumably because the adoption of a very popular book is not mainly driven by inputs received within the social network, but can be in large part driven by stimuli and mechanisms external to the online social network. As the correlations shown here correspond to static snapshots, they cannot however be used to infer causality relations between connectivity and book adoption. It is therefore also possible that the structure of the book graphs is due to the fact that people sharing the same rare book are more likely to establish social contacts than people sharing a very common book.

5.3 Spreading of books

To better understand if a user might be led to adopt a book through the influence of his/her social neighborhood, it is necessary to analyze the temporal evolution of the $G(b)$ graphs. We call $G(b,t)$ the social subgraph of users having book b at time t. $G(b,t)$ can evolve because of new users arriving in the social network who have b in their library, users leaving, or users adding/removing b to/from their library. For the purpose of detecting influence patterns, we disregard the newcomers (who might or not fill their own library with the books they have read) and users leaving the network, and focus on the graph $G^*(b,t)$ restricted to the users who are present in all the considered snapshots. Moreover, for simplicity, we neglect the (very rare) events of book deletion: once a book is adopted by a user, we assume that it is present in his/her library at any future time. In this context, we formally define the set of *adopters* of a book b between time $t-1$ and t as $A^*(b,t) = G^*(b,t) \backslash G^*(b,t-1)$.

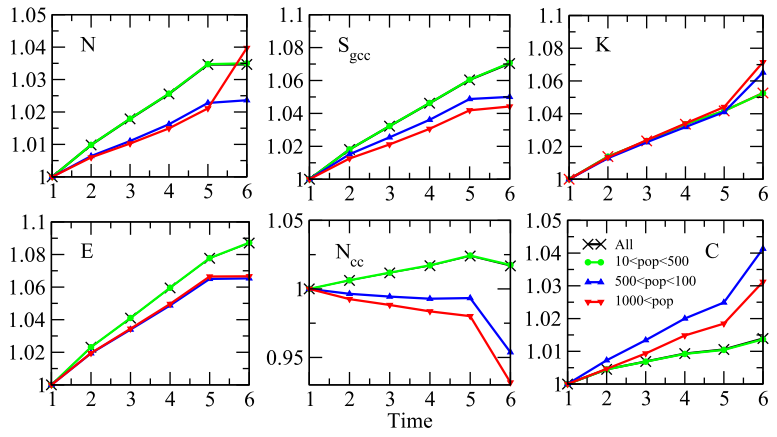

Figure 12 Temporal evolution of some properties of the book graphs. Number of nodes ($N = |A(b)|$), number of edges (E), size of the greatest connected component (S_{gcc}), number of connected components (N_{cc}), average out-degree (K) and clustering coefficient (C) are shown. All values are normalized by the initial value (at $t = 1$).

In Figure 12 the evolution of some properties of the graphs $G^*(b,t)$ is shown. Most of the values (N, E, K, C, S_{gcc}) grow in time, revealing the expansion and the increase of density and cohesion of the greatest component. The only exception is observed for the decreasing trend in the number of connected components for the graphs of the books with medium or high popularity. This can be explained by the fact that if a book is widespread over the social network it is more likely that a new adopter can create a bridge between two components of $G^*(b,t-1)$, thus reducing their number.

For every adopter, we measure the fraction of users that could potentially have played an influence in the book adoption process. If a book is adopted in the time span $[t,t+1]$, the users that may have influenced the adopter are her out-neighbors who already have that book in their library at time t.[b] We specifically focus only on the out-neighbors because users are explicitly notified of their new book adoptions, while a user may not be aware of the activity of his/her in-neighbors. Consequently, we denote the number of user u's out-neighbors at time t having book b as $K_b(u)$ and the fraction of such users over all u's out-neighbors as $F_b(u) = K_b(u)/K_{\text{out}}(u)$.

The distributions of K_b and F_b for the users u who adopt b in $[t,t+1]$ are shown in Figure 13, together with the same distributions restricted to the users u who still do not have adopted b at $t+1$. The curves for the two user categories are very different for both measures, thus revealing that users who adopt a particular book have been exposed, on average, to a higher number of users who had previously put that book in their libraries. In particular, the probability of having no out-neighbors at t with the target book in their library is much lower for the adopters (0.66) than for the non-adopters (0.98). Furthermore, as shown in Figure 14, the average K_b at fixed values of K_{out} is much higher for adopters than for non-adopters, even if a positive correlation is found in both cases, meaning that adopters have been considerably more exposed, on average, to other users exposing the adopted book. Such clear differences between the two cases of adopters and non-adopters represents a very strong evidence of the presence of an influence effect in the process of book adoption.

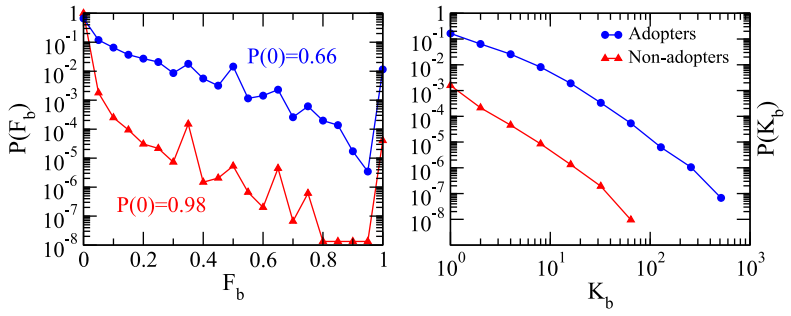

Figure 13 Compared properties of adopters and non-adopters. Distributions of portion (F_b) and number (K_b) of neighbors on the social network having book b at time t for adopters of book b between t and $t+1$ and for non-adopters. For the adopters, the same distributions computed on the interaction network only are shown.

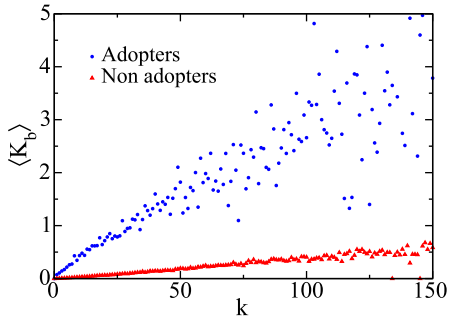

Figure 14 Compared neighborhoods of adopters and non-adopters. Average number (K_b) of out-neighbors on the social network having book b at time t, for the users having k out-neighbors. The cases of adopters and non-adopters are shown.

Interestingly, the vast majority (74%) of adopters with $F_b > 0$ exhibit values smaller than 0.2, and the average value of F_b for these adopters is rather small (0.189); on the other hand, the numbers K_b of neighbors of an adopter who already have the book b are broadly distributed. This could support two distinct hypothesis: the first one is that only a rather small number of neighbors are really influential among the neighborhood of a user; the second is that the important criterion in the adoption of a book (an 'influence threshold') is not the bare number of neighbors who have adopted a book, but the corresponding fraction among all out-neighbors, and that the influence threshold in such context is rather low.

5.4 Influence factors

As previously mentioned, users are notified of the adoption of a book by their out-neighbors: information flows in an automated way along the friendship and neighborhood links. It is thus interesting to compare the potential existence of influence effects in the book adoption process along the social links that do not support additional (non automated) communication between the users (Social\Comm) with respect to the case of social links that do (Social ∩ Comm). To this aim, we compute the probability of adoption at time t of a book b given a fixed number of neighbors who already have b at time $t-1$,

formally: $P_a(b,t|K_b)$ with $K_b = |\Gamma_{\text{out}} \cap G^*(b,t-1)|$, where Γ_{out} is the set of out-neighbors of u.

The computation of P_a for the pure social network must use out-neighbors because the information (*i.e.*, automatic notifications) flows against the direction of the edges. In the interaction network instead, both directions should be taken into account because a message sent from u to v may imply a particular interest of u in v's library or, conversely, that u is proactively suggesting a book to v. For this reason in the interaction network we consider two separate cases where K_b is computed considering the set of in-neighbors Γ_{in} or out-neighbors Γ_{out}.

Figure 15 shows the values of $P_a(b,t|K_b)$ averaged over all books and time steps, for the pure social network (Social\Comm) and the interaction network (Social ∩ Comm).

Interesting features emerge: (i) the probability of adoption is very small if $K_b = 0$ (less than $2 \cdot 10^{-4}$), and increases very rapidly as the number of out-neighbors having the considered book at $t-1$ increase; (ii) this probability tends to saturate as K_b increases above 20, showing that an additional increase in the number of out-neighbors reading the book do not increase the user's adoption probability; (iii) the probability of adoption at fixed number of out-neighbors reading the book is much larger for out-neighbors with whom an explicit communication is established; (iv) when focusing on interaction ties, receiving messages from a certain number of early adopters of a book b implies a higher probability of adoption of b than sending messages to the same number of owners of b.

The first result is a strong indication in favor of the hypothesis of effective influence between neighbors on the social graph. The second indicates that the number of influential neighbors is limited, in support of the first hypothesis outlined above. The third result supports a scenario in which direct suggestions from neighbors with whom an explicit communication exists have a stronger influencing power than the automated notification system and, in particular, the fourth result suggests that adoption is at least partially triggered by direct recommendations received by earlier adopters.

The saturation of the influence probability over a certain threshold of K_b still does not answer the long lasting question whether the adoption probability rises more with the increase of the fraction (F_b) or the number (K_b) of neighbors who are potential influencers (in this case, earlier book adopters). To give some insight into this issue, we adopt a prediction approach in which, given a book graph $G^*(b,t)$, we try to predict the shape of the

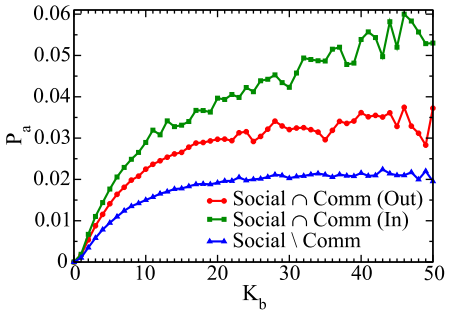

Figure 15 Book adoption probability. Probability of adoption of a book averaged on all the books and snapshots, at fixed values of the number of neighbors who have already adopted the book. Curves are depicted for the pure social network (**Social\Comm**), considering only the out-neighbors, and for the interaction network (**Social ∩ Comm**) considering the two cases of in- and out-neighbors.

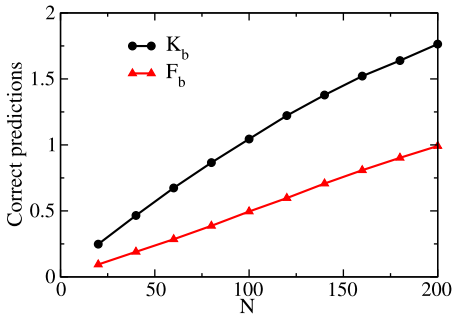

Figure 16 Prediction performance for book adoption. Prediction of new nodes being added to a book graph $G(b)$ between time t and $t+1$, based on the lists of non-adopters at time t ranked by K_b and F_b values. The number of correct prediction against the length of the prediction list is shown.

graph $G^*(b, t+1)$ based on the F_b and K_b values at time t of the users who will adopt book b at time $t+1$. In short, we rank all the users who do not have book b at time t by their F_b and K_b values, and we count how many of them in the top N entries of the rank are adopters at time $t+1$. Figure 16 shows the comparison between the metrics. For the purpose of this experiment we focus only on books with at least 20 new adopters in the time frame considered. Diffusion prediction falls out from the scope of this work, therefore we are not interested in evaluating the absolute but rather the relative performance of the two predictions. In fact, as expected, the absolute number of correctly predicted new adopters is always very low, due to the extremely high sparsity of the problem (the books spread slowly compared to the overall number of users that may potentially be influenced by their neighbors) and the simplicity of the features. However, the difference between the two curves shows clearly that K_b outperforms F_b for every value of N. This finding tends to support further the theory that influence is triggered more likely by few contacts that are able to communicate to the user and persuade him directly rather than by the portion of the social neighborhood that adopted the new item.

6 Recommending social contacts

The analysis reported in the previous sections sheds light on the dynamics of link creation in social media. Understanding the processes behind the creation of social connections allows to infer some model of network growth that can be exploited to predict the evolution of the system. In this section we will use the acquired knowledge of network dynamics to predict the creation of new links. More specifically, we propose a methodology for personalized contact recommendation that could be directly implemented on any social media like aNobii.

6.1 Prediction features

The task of predicting user pairs that will be connected in the future by a social tie can rely on two main sources of information: the structural features of the graph and the features from the user profiles. We use both types of features, considering the three main evolutionary patterns of the social graph that we previously detected.

1. *Proximity-driven link creation*. In the vast majority of cases, new neighbors are chosen among the nodes at distance 2 (*i.e.*, closing triangles) or 3 in the social graph. Restricting the analysis to pairs that reside near in the graph may miss some potential

new connections but dramatically lowers the time needed by practical algorithms for partner recommendation.

2. *Strong interaction links*. Users are influenced and inspired more by the social contacts with whom they carry out a regular communication. Taking into account the strength of the interaction links rather than (or in addition to) pure social ties could improve the prediction.
3. *Homophily-driven attachment*. Users create new connections preferentially with their most similar acquaintances. Similarity is a notion that involves all the different facets of the user profile (from geographic location to favorite books). Pairs of more similar users should therefore be considered as more likely candidates for a link creation.

A list of features that synthesizes these three principles is shown in Table 5. Most of the topological features presented have been used independently in literature for link prediction in undirected networks but can be easily adapted to the directed case. To also take into account the information concerning the weighted interaction network, we introduce a new index, the *weighted flow*, inspired by previous work on generalized degree centrality in social networks [66]. It is defined as:

$$wf(u,v) = CN(u,v) + \frac{\sum_{x \in CN(u,v)} \min(w(u,x), w(x,v))}{CN(u,v)}. \tag{2}$$

Assuming that weights on arcs denote some information flow passing between nodes, weighted flow combines the definition of common neighbors with the normalized sum of the minimum flow of information passing from the arcs connecting the two target nodes through their common neighbors. Applied to the interaction network, this metric measures both the number of potential communication channels between the two nodes and the amount of information that could have been possibly exchanged between them using their directed common neighbors as proxies.

Table 5 List of features used in the prediction of a directed link between generic users *u* and *v*, along with their description

Feature	Description	Rank
Location	Binary attribute, whether u and v belong to the same city	14
Gender	Binary attribute, whether u and v belong to the same gender	15
Age	Absolute difference of ages	12
Library	Cosine similarity between library vectors	**5**
Groups	Cosine similarity between group membership vectors	**7**
Group size	Size of the smallest group the two users have in common	**6**
Vocabulary	Cosine similarity between sets of tags used	16
Contact list	Cosine similarity of the vectors of social contacts	**2**
Outdegree	Sum of the out degrees ($k_{\mathbf{out}}(u) + k_{\mathbf{out}}(v)$)	11
Preferential attachment	Product of the out degrees ($k_{\mathbf{out}}(u) \cdot k_{\mathbf{out}}(v)$)	13
Common neighbors	Number of common neighbors, directed case ($CN(u,v) = \lvert\Gamma_{\mathbf{out}}(u) \cap \Gamma_{\mathbf{in}}(v)\rvert$)	**4**
Triangle overlap	$\frac{CN(u,v)}{\Gamma_{\mathbf{out}}(u)}$	**1**
Reciprocation	Binary attribute, whether the inverse link (v,u) is already present	**9**
Resource allocation	$\sum_{z \in (\Gamma_{\mathbf{out}}(u) \cap \Gamma_{\mathbf{in}}(u))} (\frac{1}{k_{\mathbf{out}}(z)})$ [37]	**3**
Local path	Linear combination of common neighbors and common distance-2 neighbors ($CN + \epsilon \cdot CN_2$) [37]	10
Weighted flow	$wf(u,v) = CN(u,v) + \frac{\sum_{x \in CN(u,v)} \min(w(u,x), w(x,v))}{CN(u,v)}$	**8**

$\Gamma_{\mathbf{in/out}}(u)$ denotes the set of u's in/out neighbors, $k_{\mathbf{out}}(u) = \lvert\Gamma_{\mathbf{out}}(u)\rvert$, and $w(x,y)$ is the weight of the tie between x and y. The rank reported is the result of the Chi Squared attribute selection method applied to our test set; the bold font of the rank indicates that the corresponding feature has been selected for the restricted feature set.

6.2 Classifier training and feature selection

Features can be combined through a supervised machine learning approach. A classifier properly trained on the mentioned features can determine, given any pair of nodes, if they are likely to create a social link between each other in the future. By knowing in advance the user pairs with higher connection probability, social contact recommendations can be sent to the endpoints, with the aim of notifying the two endpoints of the possibility of establishing a potentially interesting social connection that they may not have noticed otherwise or at least to speed up the linking process between them. We follow this approach and we discuss its effectiveness in a link recommendation scenario.

We choose to use a Rotation Tree classifier [67] that turned out, a posteriori, to be the best performing among all WEKA's [68] classifiers, and we train it with all the available features. The positive sample of the training set is built by about $10k$ pairs of users who reside at distance 2 on the social graph at the time of snapshot 1 and get connected before snapshot 6. The negative sample is given by as many pairs residing 2 hops away at snapshot 1 and that do not become connected. We consider only distance-2 neighbors because in the link recommendation task we will restrict our prediction to the closest non-connected pairs for computational efficiency reasons. Note that taking into account only distance-2 pairs makes the prediction task harder than selecting the non-connected pairs at random; this is due to the fact that the distribution of similarity values of pairs of users lying at distance 2 on the graph are more similar between positive and negative samples than for pairs of users taken at random (and hence farther away on the network with high probability).

As a preliminary check of the accuracy of the classifier, and in order to measure the relative predictive power of different features, we perform a 10-fold cross validation on the training set. Results for four different combinations of features are listed in Table 6; the predictive effectiveness of the features is measured through standard metrics such as the number of false positives and false negatives, accuracy, F-value, and area under the ROC. From the comparison it emerges clearly that the combination of structural and profile features leads to an appreciable improvement of the prediction quality, for all the performance indexes considered. Furthermore, aside from assessing that the combination of feature sets of different natures is good for the prediction, a more fine-grained exploration of the predictive potential of the features considered can help to exclude more redundant features, thus simplifying the decision process of the classifier and avoid overfitting. To this end, we executed the Chi Squared analysis for feature selection [69] to get a ranking of the predictive potential for all the features (see Table 5). We observe that features like vocabulary, gender, and preferential attachment have much less relevance than other features like the contact list or the library. In particular, we notice that features based on the triangle closure phenomenon are the most predictive.

Table 6 Prediction performance on the training set using the Rotation Forest classifier, 10-fold cross validation, with balanced positive and negative samples (10,000 examples)

Features	FP rate	FN rate	Accuracy	F-value	AUC
Profile	0.279	0.364	0.679	0.678	0.741
Structural	0.241	0.298	0.730	0.730	0.805
All	0.223	0.264	0.757	0.757	0.835
Restricted	0.219	0.279	0.751	0.751	0.826

Four different combinations of features are considered, the "Restricted" category includes a smaller sets of features designated as more predictive by the feature selection process.

By only using the top 9 features we verify that the prediction accuracy remains very stable and the False Positive rate is even slightly lower than with the full feature set (Table 6). We therefore retrain the classifier using the restricted feature set and use such classifier as the fundamental building block of our social contact recommender, described in the next subsection.

6.3 Contact recommendation

A contact recommendation service should be able to provide suggestions in real-time and on demand. Screening all the users that are not connected with the client requires a too high computational effort to meet this requirement. Therefore, we adopt a local search limited to the distance-2 neighborhood of the target user; among those potential contacts, the system outputs a fixed number N of suggestions.

To evaluate the effectiveness of this approach we build a test set of active users who established at least 20 new social ties between snapshots 1 and 6 with people who reside at distance 2 from them at snapshot 1. For each user u among such set, we apply our classifier to every pair $(u, v)|d(u, v) = 2$ and, from the set of pairs labeled positively by the classifier, we select N contacts to compose the recommendation list. The list is sorted according to the confidence score given by the classifier for each prediction. The number of actual ties created by the sampled users between time 1 and 6 is around $3k$, while the number of potential ties that could have been established by these users towards distance-2 neighbors is higher than $650k$. The goal of the classifier is to identify the $3k$ correct pairs among the $650k$ possible, with the lowest number of misclassifications. Such huge disproportion of positives and negatives instances determines a very high sparsity of the problem (density is less than 0.005), thus making the recommendation task particularly hard to solve with high accuracy.

Recommendation results are depicted in Figure 17, to measure the recommendation effectiveness we count the number of correctly predicted link creations, which account for the number of successful recommendations in this setting. We compare our recommender with two unsupervised techniques taking into account two single features separately, namely the number of common neighbors and the cosine similarity between libraries. In such unsupervised strategies, the recommendation lists are created by simply picking the N pairs with the highest scores for the considered metric. Results show that

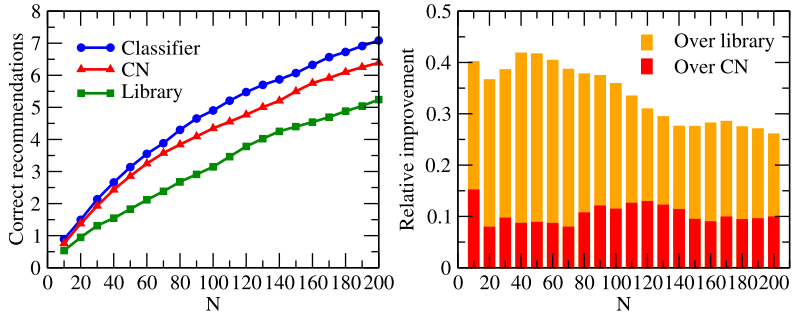

Figure 17 Recommendation results. Left: Precision at N for the recommendation made with the classifier combining all the relevant features and for two unsupervised baselines (common neighbors and library similarity). Right: Relative improvement on the classifier-based approach over the baselines.

the classifier outperforms appreciably the baselines and the number of the correctly predicted contacts grows steadily with the size of the recommendation list. However, it is surprising to observe that the relative precision (*i.e.*, the number of correct recommendations divided by the recommendation list size) is rather low, being less than 0.10 up to $N = 20$ and around 0.04 for $N = 200$.

To investigate the causes of such modest performance, we compare the obtained results with another attempt of tackling the link prediction problem from a recommendation perspective made in the Facebook social network [28]. The evaluation of the recommendation is very similar to ours with respect to the size of the network sample, the time span of the prediction and the activity of the target users. Among all the experiments that authors report, recommendation through logistic regression combining several structural graph features compares well to our approach. Nevertheless the number of correct recommendations is higher than in the aNobii case (correct recommendations at 20 is around 7.50 against ours 1.50). The main reason is due to the different sparsity of the problem. Specifically:

- In the same time span, the average number of new contacts per user in Facebook is more than six times larger than in aNobii (26 new links in Facebook *vs.* 4 in aNobii);
- The portion of new contacts residing at distance larger than 2 in aNobii is around 0.4, while in the Facebook dataset it is negligible;
- Contrary to Facebook, the aNobii network is directed and the predictions must take into account the directionality of the edge.

In Facebook, users are much more active and faster in establishing new contacts and they focus much more on their distance-2 neighbors, thus increasing the number of potential true positives over the total number of potential new contacts. Nevertheless, we underline that even in aNobii's more challenging setting the relative improvement of machine learning combination of different profile and structural features over the performance of common neighbors is comparable to the improvement obtained in the case of Facebook by previous work.

In short, the difference between the two cases can be summarized as follows. In Facebook, the decision of link creation among two people depends largely on the fact that the two endpoints have a social connection in the offline world, so that the decisional process to determine whether to add a new contact or not can be fast and simple. Conversely, in social networks with a stronger emphasis on topical interests, the items shared are more important than the personal user features (especially for neighborhood links that relate individuals who do not know each other *a priori*) and they are the main driver for the establishment of new social connections. The creation of links in such an interest network is therefore determined by the complex cognitive processes needed to relate multifaceted objects like books. This implies also a slower pace in such decisional process. Reaching definitive conclusions on this matter would require an extensive comparison between social media with different scopes (*e.g.*, music, news, photos), we believe our study can represent a contribution in this direction.

7 Discussion and conclusions

Link creation and influence are the processes on which most of the dynamics of online social media are based. In this work, we have characterized such phenomena in the case of aNobii, a network of interest for book lovers.

We have found that link formation has a strong propensity to topical and structural selection effects, reciprocity, and proximity-driven attachment. Based on these observations, we have collected a large number of both novel and state-of-the-art metrics that have a potential in predicting the formation of new links. Among such features, ranging from topical (*e.g.*, similarity between items owned by two users) to structural ones (*e.g.*, estimation of the amount of information potentially flowing from one person to the other *via* social links), we have detected the most predictive, thus shedding some light on the relative effectiveness of the main features that have been used in past work on link prediction. We have combined the best features into a classifier able to output a prediction about the future creation of a connection between any pair of nodes in the social network. We have used such classifier to produce recommendations of new social contacts for users. Differently from link prediction, that aims at predicting the global evolution of the network, link recommendation provides a contact list for every single individual and succeeds when many of the recommended contacts are actually linked by the target user. Such task is still widely unexplored and has been attempted only on general-purpose social networks with a strong accent on the user profile (*e.g.*, Facebook) rather than in interest networks like aNobii. The classifier considerably improves accuracy over simple yet very strong baselines, but the obtained performance is lower than the one reported for general-purpose online social media in previous work. The reasons for this gap likely reside in the different nature of the two cases. While in profile-focused services social aggregation is often based on the existence of a relation in the real world, that can be detected easily with simple metrics (*e.g.*, number of common friends), in interest-based networks the creation of new links is driven by cognitive processes needed to evaluate the topical interest in one profile rather than on another, that are more difficult to capture and anticipate. This finding opens the way to the exploration of the potential of prediction and recommendation in social platform with different topical focuses.

Investigation of influence complements the study on link creation. Unlike previous work, we investigate influence from both user and item perspectives. From the user side, we support with strong evidences the thesis that similarity patterns that are detected in the static network are also determined by the influence that connected users exert on each other. In particular, we observe that link creation triggers a noticeable sudden increase in the similarity between the endpoints, particularly in terms of books adopted. We inspect patterns of book adoption by modeling graphs of book spreading in time and comparing them with null models to point out their clustered and expanding nature. Based on this model, we find that the fraction of neighboring users that may have influenced an adopter is on average rather small, that the probability of adopting a book saturates as the number of neighbors already having that book increases and that the probability to adopt a book in function of the number of earlier adopters in the social neighborhood is higher if explicit communication channels exist with these neighbors. By adopting a prediction perspective, we also shed some light on the question about the fraction or the absolute amount of earlier adopter neighbors being the best indicator of higher probability of adoption, and we find that the absolute number is by far more predictive of a future adoption (even if accurate spreading prediction remains a difficult task due to the extreme sparsity of the problem and to external unobservable factors determining adoption). All these results support the idea that the 'information contagion' is a slow but relevant phenomenon in the social network and that it is usually triggered by a small number of influential users.

Another finding involves the analysis of the interaction network. For both link creation and information spreading, the interaction network has an important role in determining new connections and preferential channels of item diffusion. Many previous work showed that communication graph conveys a much stronger social signal than the pure social graph, but the implication of such stronger connections on sociological phenomena like homophily and influence had not been investigated directly before.

This work opens several natural research directions. Among possible research lines we mention the development of a model of spreading that relies on some user metadata other than the topology of the network and that could fit the phenomenon of book spreading we observed. A more thorough exploration of the possibility of predicting item spreading in contexts with slow content consumption like aNobii is also an interesting possible future extension and may open up the way to new item recommendation techniques.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors designed the research. AB, LMA, and RS performed the data analysis. LMA performed the recommendation experiments. All authors analyzed the results, wrote, reviewed and approved the manuscript.

Author details
[1]Department of Computer Science, University of Torino, Torino, Italy. [2]Centre de Physique Théorique, Aix-Marseille Université et Université du Sud Toulon Var, CNRS UMR 6207, Marseille, France. [3]Data Science Laboratory, ISI Foundation, Torino, Italy.

Acknowledgements
This work has been partially supported by the Italian Ministry for University and Research (MIUR), within the framework of the project 'Information Dynamics in Complex Data Structures' (PRIN). We acknowledge support from the Lagrange Project of the ISI Foundation supported by the CRT Foundation.

Endnotes
[a] Strictly speaking, it is impossible to prove that our crawls reached effectively the largest component. Given its characteristics and size, which are in agreement with known properties of the aNobii social system, it is however a reasonable assumption.

[b] We disregard here the possibility of interactions between users taking place outside the social network. It is clear that what can be inferred from the analysis of the online social network are only tendencies and indications, and that no absolute proof of influence effects can be obtained, as one cannot rule out effects external to the network.

Received: 1 October 2012 Accepted: 14 November 2012 Published: 5 December 2012

References
1. Mislove A, Marcon M, Gummadi KP, Druschel P, 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, pp 29-42
2. Kumar R, Novak J, Tomkins A (2006) Structure and evolution of online social networks. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '06. ACM, New York, pp 611-617. http://doi.acm.org/10.1145/1150402.1150476
3. Leskovec J, Backstrom L, Kumar R, Tomkins A (2008) Microscopic evolution of social networks. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '08. ACM, New York, pp 462-470. http://doi.acm.org/10.1145/1401890.1401948
4. Ahn YY, Han S, Kwak H, Moon S, Jeong H (2007) Analysis of topological characteristics of huge online social networking services. In: WWW '07: proceedings of the 16th international conference on World Wide Web. ACM, New York, pp 835-844
5. Aiello LM, Barrat A, Cattuto C, Ruffo G, Schifanella R (2010) Link creation and profile alignment in the aNobii social network. In: SocialCom '10: proceedings of the second IEEE international conference on social computing. IEEE Press, Minneapolis, pp 249-256
6. Mislove A, Koppula HS, Gummadi KP, Druschel P, Bhattacharjee B (2008) Growth of the Flickr social network. In: WOSN '08: proceedings of the first workshop on online social networks. ACM, New York, pp 25-30
7. Lauterbach D, Truong H, Shah T, Adamic L (2009) Surfing a web of trust: reputation and reciprocity on CouchSurfing.com. In: Computational science and engineering, IEEE international conference on, vol 4, pp 346-353
8. Weng J, Lim EP, Jiang J, He Q (2010) TwitterRank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on web search and data mining, WSDM '10. ACM, New York, pp 261-270. http://doi.acm.org/10.1145/1718487.1718520

9. Aiello LM, Barrat A, Schifanella R, Cattuto C, Markines B, Menczer F (2012) Friendship prediction and homophily in social media. ACM Trans Web 6:9
10. Chun H, Kwak H, Eom YH, Ahn YY, Moon S, Jeong H (2008) Comparison of online social relations in volume vs. interaction: a case study of cyworld. In: Proceedings of the 8th ACM SIGCOMM conference on Internet measurement, IMC '08. ACM, New York, pp 57-70. http://doi.acm.org/10.1145/1452520.1452528
11. Wilson C, Boe B, Sala A, Puttaswamy KP, Zhao BY (2009) User interactions in social networks and their implications. In: Proceedings of the 4th ACM European conference on computer systems, EuroSys '09. ACM, New York, pp 205-218. http://doi.acm.org/10.1145/1519065.1519089
12. Leskovec J, Horvitz E (2008) Planetary-scale views on a large instant-messaging network. In: Proceedings of the 17th international conference on World Wide Web, WWW '08. ACM, New York, pp 915-924. http://doi.acm.org/10.1145/1367497.1367620
13. Viswanath B, Mislove A, Cha M, Gummadi KP (2009) On the evolution of user interaction in Facebook. In: WOSN '09: proceedings of the 2nd ACM workshop on online social networks. ACM, New York, pp 37-42
14. Benevenuto F, Rodrigues T, Cha M, Almeida V (2009) Characterizing user behavior in online social networks. In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, IMC '09. ACM, New York, pp 49-62. http://doi.acm.org/10.1145/1644893.1644900
15. Gonçalves B, Perra N, Vespignani A (2011) Modeling users' activity on Twitter networks: validation of Dunbar's number. PLoS ONE 6(8):e22656. http://dx.doi.org/10.1371%2Fjournal.pone.0022656
16. Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '03. ACM, New York, pp 137-146. http://doi.acm.org/10.1145/956750.956769
17. Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the Flickr social network. In: Proceedings of the 18th international conference on World Wide Web, WWW '09. ACM, New York, pp 721-730. http://doi.acm.org/10.1145/1526709.1526806
18. Ye S, Wu SF (2010) Measuring message propagation and social influence on Twitter.com. In: Proceedings of the second international conference on social informatics, SocInfo '10. Springer, Berlin, pp 216-231. http://portal.acm.org/citation.cfm?id=1929326.1929342
19. Cha M, Haddadi H, Benevenuto F, Gummadi KP (2010) Measuring user influence in Twitter: the million follower fallacy. In: ICSWM '10: proceedings of the 4th international AAAI conference on weblogs and social media
20. Anagnostopoulos A, Kumar R, Mahdian M (2008) Influence and correlation in social networks. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '08. ACM, New York, pp 7-15. http://doi.acm.org/10.1145/1401890.1401897
21. Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S (2008) Feedback effects between similarity and social influence in online communities. In: KDD '08: proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 160-168
22. Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. In: Proceedings of the 21st international conference on World Wide Web, WWW '12. ACM, New York, pp 519-528. http://doi.acm.org/10.1145/2187836.2187907
23. Yang J, Leskovec J (2010) Modeling information diffusion in implicit networks. In: Proceedings of the 2010 IEEE international conference on data mining, ICDM '10. IEEE Computer Society, Washington, pp 599-608. http://dx.doi.org/10.1109/ICDM.2010.22
24. Gomez Rodriguez M, Leskovec J, Krause A (2010) Inferring networks of diffusion and influence. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '10. ACM, New York, pp 1019-1028. http://doi.acm.org/10.1145/1835804.1835933
25. Au Yeung CM, Iwata T (2010) Capturing implicit user influence in online social sharing. In: Proceedings of the 21st ACM conference on hypertext and hypermedia, HT '10. ACM, New York, pp 245-254. http://doi.acm.org/10.1145/1810617.1810662
26. Getoor L, Diehl CP (2005) Link mining: a survey. ACM SIGKDD Explor Newsl 7(2):3-12
27. Cooke RJE (2006) Link prediction and link detection in sequences of large social networks using temporal and local metrics. Master thesis, Department of Computer Science, University of Cape Town
28. Backstrom L, Leskovec J (2011) Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the fourth ACM international conference on web search and data mining, WSDM '11. ACM, New York, pp 635-644. http://doi.acm.org/10.1145/1935826.1935914
29. Liben-Nowell D, Kleinberg J (2003) The link prediction problem for social networks. In: CIKM '03: proceedings of the twelfth international conference on information and knowledge management. ACM, New York, pp 556-559
30. Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am Soc Inf Sci Technol 58(7):1019-1031
31. Pavlov M, Ichise R (2007) Finding experts by link prediction in co-authorship networks. In: FEWS2007: proceedings of the workshop on finding experts on the web with semantics at ISWC/ASWC2007, Busan, South Korea
32. Popescul A, Popescul R, Ungar LH (2003) Structural logistic regression for link analysis. In: Proceedings of the second international workshop on multirelational data mining
33. Hasan MA, Chaoji V, Salem S, Zaki M (2006) Link prediction using supervised learning. In: Proceedings of SDM '06 workshop on link analysis, counterterrorism and security
34. O'Madadhain J, Hutchins J, Smyth P (2005) Prediction and ranking algorithms for event-based network data. ACM SIGKDD Explor Newsl 7(2):23-30
35. Zheleva E, Getoor L, Golbeck J, Kuter U (2008) Using friendship ties and family circles for link prediction (poster paper). In: 2nd SNA-KDD workshop on social network mining and analysis. ACM, Las Vegas
36. Sachan M, Ichise R (2010) Using semantic information to improve link prediction results in networked datasets. Int J Eng Technol 2(4):334-339
37. Zhou T, Lü L, Zhang YC (2009) Predicting missing links via local information. Eur Phys J B 71(4):623-630. Special issue: The physics approach to risk: agent-based models and networks
38. Lü L, Ci-Hang J, Zhou T (2009) Effective and efficient similarity index for link prediction of complex networks. arXiv:0905.3558

39. Song HH, Cho TW, Dave V, Zhang Y, Qiu L (2009) Scalable proximity estimation and link prediction in online social networks. In: IMC '09: proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, New York, pp 322-335
40. Schifanella R, Barrat A, Cattuto C, Markines B, Menczer F (2010) Folks in folksonomies: social link prediction from shared metadata. In: WSDM '10: proceedings of the third ACM international conference on web search and data mining. ACM, New York, pp 271-280
41. Leroy V, Cambazoglu BB, Bonchi F (2010) Cold start link prediction. In: SIGKDD '10: proceedings of the 16th ACM conference on knowledge discovery and data mining. ACM, Washington
42. Bilgic M, Namata GM, Getoor L (2007) Combining collective classification and link prediction. In: ICDMW '07: proceedings of the seventh IEEE international conference on data mining workshops. IEEE Computer Society, Washington, pp 381-386
43. Tylenda T, Angelova R, Bedathur S (2009) Towards time-aware link prediction in evolving social networks. In: SNA-KDD '09: proceedings of the 3rd workshop on social network mining and analysis. ACM, New York, pp 1-10
44. Dunlavy DM, Kolda GK, Acar E (2010) Temporal link prediction using matrix and tensor factorizations. arXiv:1005.4006
45. Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabási AL (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci USA 104(18):7332-7336. http://dx.doi.org/10.1073/pnas.0610245104
46. Lü L, Zhou T (2009) Role of weak ties in link prediction of complex networks. In: CNIKM '09: proceedings of the 1st ACM international workshop on complex networks meet information and knowledge management. ACM, New York, pp 55-58
47. Gilbert E, Karahalios K (2009) Predicting tie strength with social media. In: Proceedings of the 27th international conference on human factors in computing systems, CHI '09. ACM, New York, pp 211-220. http://doi.acm.org/10.1145/1518701.1518736
48. Benchettara N, Kanawati R, 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, pp 326-330
49. Kunegis J, De Luca E, Albayrak S (2010) The link prediction problem in bipartite networks. In: Hullermeier E, Kruse R, Hoffmann F (eds) Computational intelligence for knowledge-based systems design. Lecture notes in computer science, vol 6178. Springer, Berlin, pp 380-389
50. Leskovec J, Huttenlocher D, Kleinberg J (2010) Predicting positive and negative links in online social networks. In: WWW '10: proceedings of the 19th international conference on World Wide Web. ACM, New York, pp 641-650
51. Taskar B, Wong MF, Abbeel P, Koller D (2003) Link prediction in relational data. In: NIPS '03: neural information processing systems conference, Vancouver, Canada
52. Getoor L, Friedman N, Koller D, Taskar B (2003) Learning probabilistic models of link structure. J Mach Learn Res 3:679-707
53. Freeman LC (1979) Centrality in social networks: conceptual clarification. Soc Netw 1(3):215-239. http://dx.doi.org/10.1016/0378-8733(78)90021-7
54. Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge
55. Bollobas B (1998) Modern graph theory. Springer, Berlin
56. Marlow C, Naaman M, Boyd D, Davis M (2006) HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: HYPERTEXT '06: proceedings of the seventeenth conference on hypertext and hypermedia. ACM, New York, pp 31-40
57. Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89:208701
58. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47-97
59. Newman MEJ (2001) Clustering and preferential attachment in growing networks. Phys Rev E 64(2):025102
60. Kleinberg JM, Kumar R, Raghavan P, Rajagopalan S, Tomkins AS (1999) The web as a graph: measurements, models and methods. In: Computing and combinatorics. Lecture notes in computer science, vol 1627, pp 1-18
61. Kumar R, Raghavan P, Rajagopalan S, Sivakumar D, Tomkins A, Upfal E (2000) Stochastic models for the web graph. In: Proceedings of the 41th IEEE symposium on foundations of computer science (FOCS), pp 57-65
62. Liben-Nowell D, Novak J, Kumar R, Raghavan P, Tomkins A (2005) Geographic routing in social networks. Proc Natl Acad Sci USA 102(33):11623-11628
63. Lee C, Scherngell T, Barber MJ (2009) Real-world separation effects in an online social network. Technical report. http://arxiv.org/abs/0911.1229
64. McPherson M, Lovin LS, Cook JM (2001) Birds of a feather: homophily in social networks. Annu Rev Sociol 27:415-444. http://dx.doi.org/10.1146/annurev.soc.27.1.415
65. Barrat A, Barthlemy M, Vespignani A (2008) Dynamical processes on complex networks, 1st edn. Cambridge University Press, New York
66. Opsahl T, Agneessens F, Skvoretz J (2010) Node centrality in weighted networks: generalizing degree and shortest paths. Soc Netw 32(3):245-251
67. Rodriguez JJ, Kuncheva LI, Alonso CJ (2006) Rotation forest: a new classifier ensemble method. IEEE Trans Pattern Anal Mach Intell 28(10):1619-1630
68. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. ACM SIGKDD Explor Newsl 11:10-18
69. Liu H, Setiono R (1995) Chi2: feature selection and discretization of numeric attributes. In: Proceedings of the seventh international conference on tools with artificial intelligence, TAI '95. IEEE Computer Society, Washington, pp 388-391

doi:10.1140/epjds12
Cite this article as: Aiello et al.: **Link creation and information spreading over social and communication ties in an interest-based online social network.** *EPJ Data Science* 2012 **1**:12.