Vidmer et al. *EPJ Data Science* (2015) 4:20
DOI 10.1140/epjds/s13688-015-0057-x

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Unbiased metrics of friends' influence in multi-level networks

Alexandre Vidmer*, Matúš Medo and Yi-Cheng Zhang

*Correspondence: alexandre.vidmer@unifr.ch
Department of Physics, University of Fribourg, Chemin du Musée 3, Fribourg, 1700, Switzerland

Abstract

The spreading of information is of crucial importance for the modern information society. While we still receive information from mass media and other non-personalized sources, online social networks and influence of friends have become important personalized sources of information. This calls for metrics to measure the influence of users on the behavior of their friends. We demonstrate that the currently existing metrics of friends' influence are biased by the presence of highly popular items in the data, and as a result can lead to an illusion of friends influence where there is none. We correct for this bias and develop three metrics that allow to distinguish the influence of friends from the effects of item popularity, and apply the metrics on real datasets. We use a simple network model based on the influence of friends and preferential attachment to illustrate the performance of our metrics at different levels of friends' influence.

Keywords: social influence; multi-level network; social metrics; friendship network

1 Introduction

The use of friends' influence in the spreading of information has become an important mean of advertisement by companies on the Internet [1]. This type of advertisement is strongly mediated by the relations between users in the network and resembles the spreading of infectious diseases [2]; it is thus referred to as *viral marketing*. One of the most famous example is Google's e-mail account *Gmail*, which was only accessible via an invitation from another user in its early days.

It is thus important to be able to measure and understand the underlying mechanisms of the social spreading of content. The influence of users sending purchase recommendations to their friends has been studied in [3]. This study shows that this type of explicit viral marketing contributes only marginally to the total sales, and that users can become resilient when they are exposed to too many recommendations from friends or simply too many of their friends' actions. It is shown in [4] that users with high numbers of friends need more exposure of their friends' actions before doing the same actions themselves (for example watching a video or buying an item), thus leading to an apparent resilience effect. The authors correct this effect by taking into account the visibility of the actions of friends to compute the exposure, as the visibility of friends' actions depends on the way the website displays it to its users. Another aspect of the social interactions is the homophily, which should not be confounded with influence of friends [5]. In the latter case, a user consumes a good because of the behavior of another user, while in the case of homophily the users happen to consume the same good because they have similar tastes. These two effects are distinguished using statistical matching of users in [6, 7]. In [8], the influence of friends is distinguished from homophily by randomizing the timing of users' actions.

Many efforts have been invested in modeling the spreading of information through social bonds. One of the main diffusion models in marketing and management is the Bass model [9], which is generally used to study the rates of innovation and imitation in the spreading of ideas. The spreading of information is treated similarly the spreading of an infection in the following models; a user is thus said to be infected when he gets influenced by his friends. In a more recent model, the threshold model [10], each user is endowed with a personal threshold value and becomes infected when the number of infected friends meets this threshold. A generalization of this model, named the *linear threshold model*, uses the ratio of infected friends and/or links weights between friends [11, 12]. In the cascade model [13], a user has a certain probability to become infected each time one of his neighbors becomes infected.

However, most of those models only account for the influence occurring inside the network, excluding the possibility of external influence. It may be difficult to distinguish the influence of friends from the one due to external influence. A recent work takes into account both means of infection by modeling the internal and external influences and combining them into an exposure function (i.e. the strength at which users propagate an information) [14]. A model in which each user has its own time-dependent influence function has been developed in [15] and shows that the influence function depends on the type of contents. The Recommendation of music in Last.fm has been successfully improved by temporally separating the influence of friends from the general influence [16]. The evolution of both the users' friendship relations as well as the users' group affiliations have been modeled in [17]. Models of propagation can also help to recover the social ties behind the adoption patterns [18].

Social spreading is not the only mechanism at work in networks. In many empirical systems, such as the scientific collaboration networks [19], the metabolic networks [20], and the social networks [21], the number of interactions attached with individual nodes follows a broad distribution which is often, though not always, of a power-law kind [22]. A simple model based on the popular *rich get richer* principle was successfully used to reproduce the power-law distribution in the *WWW* (World Wide Web) [23]. However, this model predicts a strong correlation between item popularity and item age that is not observed in the WWW [24], nor in the citation network [25]. To solve this shortcoming, the preferential attachment model was refined by including a decaying time factor and a relevance score for nodes [26].

In this work, the input data are represented as a multi-level network which consists of a monopartite social network of users and a bipartite user-item network where two distinct type of nodes, user and item nodes, are present. Links in the social network represent social connections/friendship relations between the users. Links in the user-item network represent interactions between the users and the items (depending on the system, the interactions can correspond to collecting, buying, reviewing, or otherwise connecting with an item). We develop three new metrics to measure three different aspects of friends' influence in these networks. Contrary to research that attempts to find the most influential users in complex networks [27–29] that often uses only the user social network, we use detailed time information on user actions and aim at measuring the strength of user influence over individual users. With the first metric we study the probability that a user collects an item depending on his friends behavior. We show that with an appropriate rescaling of the number of friends who have collected an item, we obtain a metric which is sensitive to the influence of friends, as opposed to the raw unrescaled number which is basically unaffected by friends' influence. The second metric studies the influence between pairs of friends by comparing the number of times a user influences one of his friends with the number of times this would happen in a null model. This metric is particularly useful to detect the influence between moderately active users who actually constitute the majority of the users due to the free-scale nature of the network. The third metric measures the spreading of individual items in the friendship network of users by comparing the original network with a subnetwork comprising only the users who collected a specific item. This metric is sensitive to niche items which are mainly popular in small groups of friends.

We apply the three metrics on data from Yelp which is a website where users write reviews on real world businesses. In order to test the ability of our metrics to detect the influence of friends, we apply the metrics to randomized Yelp data and show that, unlike the existing metrics such as exposure and contagion, they exhibit significantly different patterns compared to those found in the original data. In addition, we apply the metrics on artificial networks grown with preferential attachment mixed with the influence of friends, and show that the new metrics are able to distinguish among networks created with different amounts of friends' influence. Finally, we show that our metrics perform well also on data from the social news website Digg.com.

2 Datasets

The number of users and items are denoted as U and I, respectively. Throughout this paper, we use Latin letters i and j to label the users and Greek letters α and β to label the items. The number of items collected by user i is the degree k_i of user i in the bipartite network, and the number of users who have collected item α is the degree k_α of the item. The number of friends f_i of user i is the degree of this user in the social network.

Two different real datasets are used in this work. The first one is the round 4 of the Yelp academic challenge dataset [30]. Yelp is a website where users can review and rate various businesses such as restaurants, doctors, and bars. This dataset is particularly suitable for our study because it features both a social component - users can explicitly select other users as friends - and a bipartite component - users can review businesses and give them scores in the integer rating scale from 1 to 5. Furthermore, rating time stamps are available which makes it possible to detect who followed whom in their collection patterns.

Based on the data, we build two networks. The first one is the friendship network, in which users are represented by nodes, and links connect the users who are marked as friends in the data. The second one is bipartite, there is a link between a user and a business if the user has reviewed the business independently of the actual ratings given by the user to the business.

The original dataset contains 252,898 users, 42,153 businesses, 955,999 friendship links, and 1,125,458 reviews. For our analysis, we keep only the users who have at least one friend and at least one review. Similarly, we consider only the businesses that received at least one review. The resulting dataset contains 123,368 users, 41,958 businesses, 955,999 friendship links and 804,789 review links.

The second dataset is composed by data from Digg.com. On this website, every user can post a news (an item) and other users can then vote for it ('digg' it). We consider that a user has collected a news if he has voted for it. A news is first only visible to the friends of the user who published it and thus it propagates only by social spreading. Once the news has earned enough votes, it is promoted to the Digg front page where every user can see it. Our dataset contains only news that were promoted to the front page. The life cycle of items is very different for Digg and Yelp networks, which is natural because news are only interesting for a short time after their publication, while a restaurant can remain popular for a long period of time. A business gets 90% of its number of reviews in one thousand days, while a news gets 90% of its votes fifteen hours after its publication. The dataset that we use here was obtained from [31]. After applying the same cleaning procedure as we did for the Yelp data, we are left with 123,368 users, 3,553 items, 2,337,418 user-item links and 874,745 friendship links.

3 Metrics of friends' influence

The most straightforward and quite common approach to quantify influence of friends in real data is based on measuring the probability that a user i collects an item α, given that $f_{i\alpha}$ of i's friends have already collected it [3, 32–36]. In the rest of the paper, we refer to $f_{i\alpha}$ as the *exposure* of user i to item α (assuming that a user is exposed to an item whenever one of his friends collects it). However, this measurement is strongly influenced by the heterogeneity of item popularity. Indeed, if we randomly distribute an item α to k_α different users, each user collects item α with probability k_α/U. As a result, a User i with f_i friends will thus have on average $f_i k_\alpha/U$ friends that have collected the item. In a random model, the expected number of friends that have collected an item is directly proportional to the degree k_α of the item. In most online networks, the growth is driven by preferential attachment which implies that the likelihood for an item to receive a new link is proportional item degree k_α [23]. Because of shared proportionality to item degree, if we plot the probability for a user to collect an item as a function of the number of friends that have collected it $f_{i\alpha}$, we essentially see the effect of preferential attachment. In order to differentiate the influence of friends from preferential attachment, we propose the normalized exposure in the form

$$n_{i\alpha} = \max_{t<t_{i\alpha}} \left(\frac{f_{i\alpha}(t)}{f_i} \frac{k_{\min}}{k_\alpha(t)} \right), \tag{1}$$

where $t_{i\alpha}$ is the time at which user i collects item α. The ratio $f_{i\alpha}(t)/f_i$ complies the observation that users with a larger number of friends need more friends who collect an item before they tend to collect it themselves [4]. We take the maximum value reached by the quantity in parenthesis over time so that we measure the largest signal sent to the user. Indeed, the ratio in the parentheses is not strictly growing because if the item becomes more popular ($k_\alpha(t)$ increases) but no new friends collect it ($f_{i\alpha}(t)$ remains constant), the quantity decreases. Due to the normalization, the values of the newly suggested normalized metrics are uncertain for little popular items and users with a low number friends. To limit the influence of these noisy estimates, we consider only items with current degree $k_{\min}$ or more and users with $f_{\min}$ friends or more ($k_{\min}$ and $f_{\min}$ are set to 10 and 5, respectively). The factor $k_{\min}$ is introduced in Eq. (1) in order to make the upper bound of $n_{i\alpha}$ independent of $k_{\min}$. The choice of $k_{\min}$ and $f_{\min}$ does not alter the results qualitatively. We compute $n_{i\alpha}$ for all user-item pairs. We then compute for n between 0 and 1 the ratio of the number of $n_{i\alpha}$ pairs greater or equal to n where user i has eventually collected item α over the number of all $n_{i\alpha}$ pairs greater or equal to n. The result is the cumulative probability that a user i collects an item α given that $n_{i\alpha}$ is greater than n.

If a user collects an item before one of their friends, we say that there is a *contagion* between the user and the friend. We introduce n_{ij} as the number of items first collected by user i that are later collected by user j (note that $n_{ij} \neq n_{ji}$ in general). This quantity was used in previous studies to quantify the social contagion occurring in social systems [37, 38]. A version normalized by the degree of the infected user was also used prior to this work [39]. However, if two friends are very active, it is likely that they have a larger number of items in common than less active users. To remove this bias, we normalize n_{ij} with the expected number of items that user i and j would have in common in a random case. Assuming $k_i, k_j \ll I$, user i selecting items randomly has the probability k_j/I to collect an item that user j has previously collected. As a result, the expected number of items that the users share is k_ik_j/I. We now introduce the *normalized contagion count* between two users as

$$c_{ij} = \frac{n_{ij}}{k_ik_j} k_{\min}, \tag{2}$$

where n_{ij} is the number of items collected first by user i and later by user j. When computing this score, we again take into account only the users who have collected at least $k_{\min}$ items, which helps to reduce the noise. The factor $k_{\min}$ is introduced in Eq. (2) to rescale its values and assure that the maximum value of c_{ij} is 1 independently of the value of $k_{\min}$. The maximal value of c_{ij} is $k_{\min}/k_i$ when $k_j = n_{ij}$, thus favoring the little active users who can in general reach higher c_{ij} values than highly active users. Introducing the threshold $k_{\min}$ again serves to limit the level of noise and its chosen value does not alter the results qualitatively. Due to the free-scale distribution of users activity in online networks, the low-active users are a very important fraction of the network and should not be neglected.

With the last metric, we are interested in measuring the topological features of item spreading. If social influence through friendship links is significant, we expect that the items spread more densely between friends than among users with no explicit relationships. To measure this effect, we introduce the *spreading concentration* r_α. For each user who has collected item α, we compute the ratio between the number of friends who have collected item α over the total number of friends (see Figure 1 for an illustration) and average the result over all these users. The score can be written as

$$r_\alpha = \frac{1}{k_\alpha} \sum_{i \in \mathcal{N}(\alpha)} \frac{\sum_{j \in \mathcal{F}(i)} a_{j\alpha}}{f_i}, \tag{3}$$

where $\mathcal{N}(\alpha)$ is the set of user who have collected item α, $\mathcal{F}(i)$ is the set of users who are friends with user i, and $a_{j\alpha}$ is an element of the adjacency matrix of the bipartite network, which is 1 if user j has collected item α and 0 otherwise. Contrary to the two previously introduced metrics, to our knowledge there is no equivalent of this metric in the existing literature. When r_α is zero, item α has been collected by users who are not connected with each other, which indicates that influence of friends played a negligible role in the way how the item has spread in the society. When r_α is high (close to one), item α has been collected

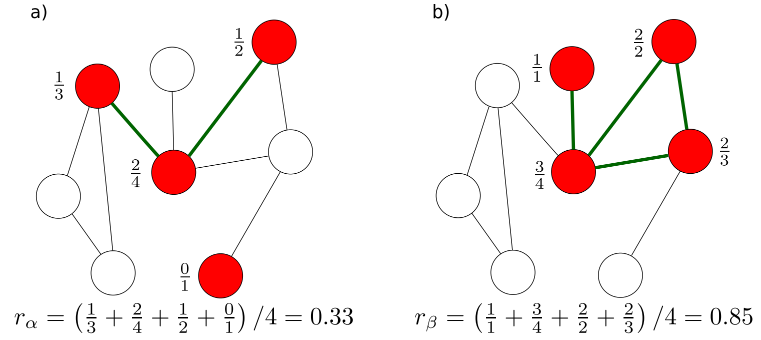
a)
b)
$\frac{1}{3}$
$\frac{1}{2}$
$\frac{2}{4}$
$\frac{0}{1}$
$\frac{1}{1}$
$\frac{2}{2}$
$\frac{3}{4}$
$\frac{2}{3}$
$r_\alpha = \left(\frac{1}{3} + \frac{2}{4} + \frac{1}{2} + \frac{0}{1}\right)/4 = 0.33$
$r_\beta = \left(\frac{1}{1} + \frac{3}{4} + \frac{2}{2} + \frac{2}{3}\right)/4 = 0.85$

Figure 1 Illustration of the spreading concentration. Spreading concentration for an item α which rarely spreads through friendship links **(a)** and item β which often spreads through friendship links **(b)**. Nodes and links represent users and their friendship relations, respectively. A node is colored when the corresponding user has collected a given item. A link is highlighted if it connects two users who have both collected a given item.

by tightly connected users, which suggests that social relations significantly contributed to the item's spreading among the users.

4 Randomization of the input data by rewiring

A metric of social influence produce some signal even when social influence plays no role in the evolution of the analyzed system. To assess whether the observed signal is in fact an indication for influence of friends, we compare the metrics applied to the original data with the metrics applied on two randomized versions of the data. To randomize the data, we apply a simple reshuffling procedure to either the user-item or the user-user network. The procedure is as follows: for each link of the selected network, we choose another link at random and we exchange one of the end nodes between them (in the bipartite network, links i-α and j-β are rewired to i-β and j-α, respectively with $i \neq j$, and we proceed similarly for the friends network). In the user-item network, the time of the action is kept on the item side. Note that this procedure keeps the degree values of both user and item nodes unchanged. By reshuffling the social network, the preferences of the users are maintained but they are assigned with random friends. While the resulting network features no influence of friends, some friends may still have similar preferences and thus have a number of items in common. By reshuffling the user-item network, we obtain a network which features no user preferences at all (everyone's collection becomes random) and consequently no influence of friends. While the two rewiring processes differ, they both remove the social influence of the data and thus can constitute a good benchmark to evaluate whether the different metrics are able to distinguish influence of friends from other effects. Instead of the rewiring procedure, one could use the configuration model [40] to produce randomized networks. Upon the same constraints (no multi-links in the bipartite network and no self-loops in the social network), the results obtained in the two ways are equivalent.

5 Artificial model

To complement the binary comparison with rewired data (some influence of friends vs zero influence of friends), we now introduce a network model where the contribution of friends' influence can be varied continuously. In this model, every user is subject to two different types of influence: from friends and global from the outside world. The global influence is modeled using preferential attachment with decaying relevance which has been shown to well represent the citation patterns of scientific papers [26] and the dynamics of content popularity in online networks [41]. The influence of friends is assumed to be proportional to the number of friends that have collected an item. In order to focus on the growing bipartite network, we do not attempt here to model the process of friendship creation and directly use the friendship network from Yelp. This allows an easier comparison of the artificial bipartite network with the real one. We grow the bipartite network by adding one item in each time step until the final number of items is equal to that in the real data. At each time step, m links are added to the network where m is chosen in such a way that the final number of links is equal to that in the real data ($m = 19$ for the Yelp data), which allows us to directly compare the structure of the artificial network with that of the real data.

When adding a link, the probability to choose user i is proportional to k_i^Y where k_i^Y is the final number of reviews of users i in the Yelp network. This makes the resulting distribution of user degree similar to that in the real data (see Figure 2(a)). To choose an item for the given user, with probability p we use the decaying relevance model (which models global influence). Otherwise (i.e., with probability $1 - p$), we use the influence of friends model. The probability to choose item α in the two respective cases is

$$P^{\text{glob}}(\alpha, t) = \frac{k_\alpha(t) R_\alpha(t)}{\sum_\beta k_\beta(t) R_\beta(t)} \quad \text{with probability } p, \tag{4}$$

$$P^{\text{soc}}(\alpha, t) = \frac{f_\alpha^i(t)}{\sum_\beta N_\beta^i(t)} \quad \text{with probability } 1 - p. \tag{5}$$

Here $f_\alpha^i(t)$ is the number of friends of user i that have collected item α at time t, $k_\alpha(t)$ is the number of links connected to item α at time t, $R_\alpha(t) = R_\alpha \exp[-\beta(t - t_\alpha)]$ is the current relevance of item α, and R_α is the initial relevance of item alpha which is drawn from the exponential distribution $\lambda \exp^{\lambda x}$, with x restricted to the range $[1, \infty)$. While a small value of λ leads to a heterogeneous distribution of initial item relevance, a high value results in a narrow distribution of relevance and thus a narrow distribution of item degree. The small λ setting leads to a power-law distribution of final item popularity [26], which is similar to the distributions found in various real systems [22]. We choose the time decaying parameter β in order to maximize the agreement between the real and model item degree

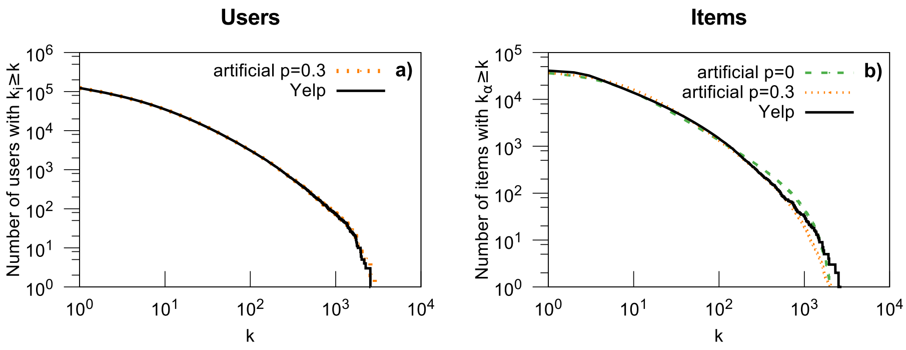

Figure 2 Degree distribution. Panels show the degree distribution of users and items in the users-item network, respectively, for both real and artificial networks.

distribution (see Figure 2 for the result with $\beta = 0.025$ and $\lambda = 1.5$). When the time decay is fast (β is large) the relevance of items quickly tends to zero for every item. When the time decay is slow (β is small but non-zero), the items with large initial relevance have long time to attract attention, resulting in high item degree and a broad item degree distribution.

6 Results on the Yelp and model data

We now study the behavior of all the described influence of friends metrics on both the Yelp data and the model data.

6.1 Exposure and normalized exposure

In Figure 3(a) we present the probability that a user collects an item as a function of the number of friends who have collected the item. Similarly to [3], we find that the quick initial increase is followed by a saturation. While the authors of [3] claim that this behavior is due to the users perceiving too many recommendations of an item from their friends as spam, we argue that this is not the case: the two randomized networks show the same pattern despite the absence of friends' influence in the randomized data. Furthermore, the same is true for all curves obtained on the model data (see Figure 3(b)) regardless of the strength of friends influence (which is controlled by the parameter p) where there is no built-in aversion to items shared by a high number of friends. Note also that the curves corresponding to different values of p in Figure 3(b) largely overlap (except for $p = 0.5$) which means that this measurement does not allow us to distinguish systems with very different levels of friends' influence. Based on these observations and our previous

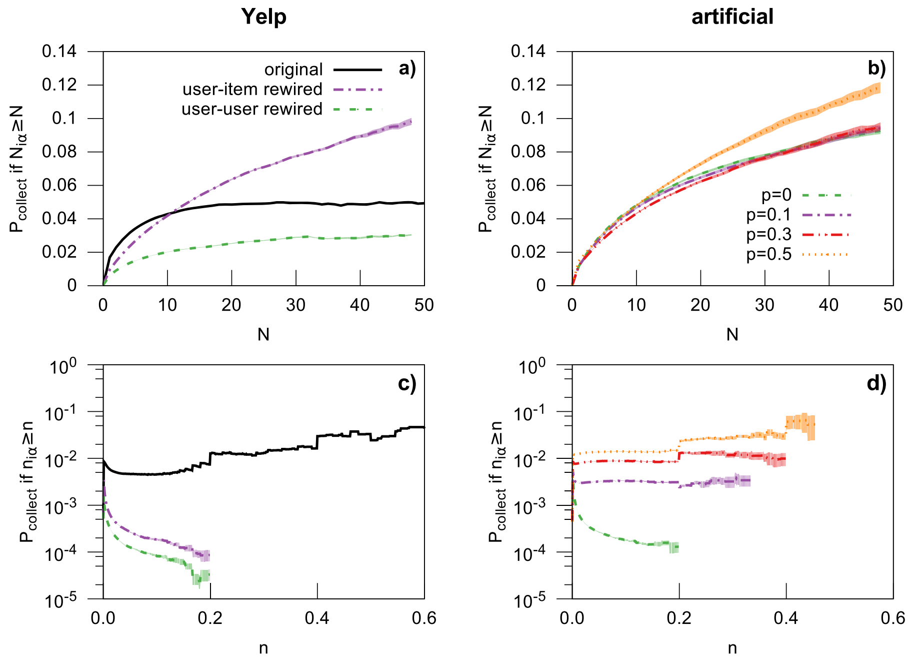

Figure 3 Normalized exposure in the Yelp data. Panels **(a)** and **(b)** show the probability that a user collects an item when its exposure (the number of the user's friends who have already collected it) is at least *N* for the Yelp and artificial data, respectively. Panels **(c)** and **(d)** show the probability that a user collects an item when its normalized exposure is at least *n*. Here (and in the following figures) shaded areas indicate the standard error on the mean computed over ten realizations.

discussion in Section 2, we argue that the heterogeneity of item popularity distorts the exposure metric to such an extent that it is not a reliable measure of friends' influence. The results are very different for the proposed normalized exposure metric. As shown in Figure 3(c), the collection probability increases with the normalized exposure in the original data and it decreases steeply (note the logarithmic scale on the vertical axis) in the randomized data, thus allowing for a clear distinction between systems with and without influence of friends. The same is true for the model data in Figure 3(d) where the curves corresponding to different values of p do not overlap (as opposed to Figure 3(b)).

6.2 Contagion count and normalized contagion count

Since the contagion count focuses on co-collection patterns among individuals, random rewiring of the user-item network is more appropriate than rewiring of the user-user network. The reason for this is that while the former effectively eliminates personal taste of users, the latter maintains them. When two users with similar taste are connected in the rewired user-user network, they can still achieve a significant contagion count despite the absence of a true friends influence between them. Figure 4(a) shows that the contagion count is able to distinguish between the original and the randomized network. However, the metric performs worse on the artificial data shown in Figure 4(b) where it fails to distinguish the data corresponding to different values of p (the only noticeable difference happens when half of the content is propagated by influence of friends). The normalized contagion count is shown for the Yelp and model data in Figure 4(c) and (d), respectively. While the difference between the original network and the reshuffled ones looks similar to the unnormalized contagion count, in the artificial networks we can easily distinguish between datasets produced with different values of the social influence parameter p.

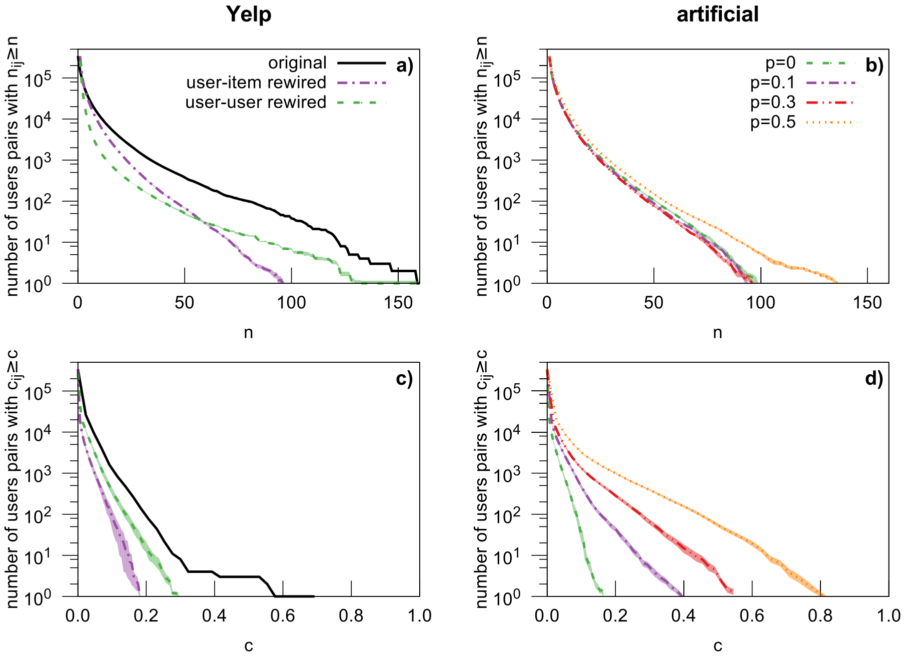

Figure 4 Normalized contagion count in the Yelp data. Panels **(a)** and **(b)** show the distribution of the raw contagion count for the Yelp and model data, respectively. Panels **(c)** and **(d)** do the same for the normalized contagion count.

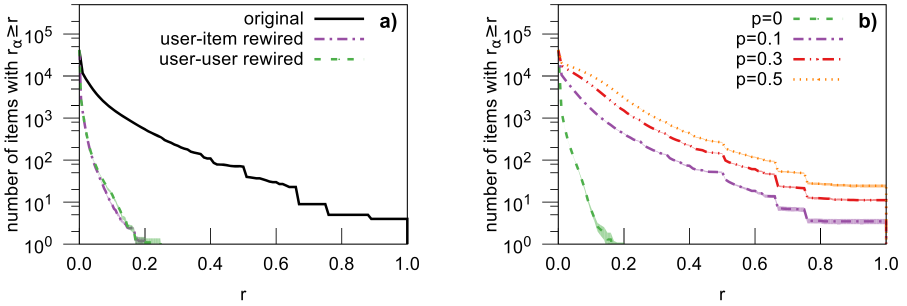

Figure 5 Spreading concentration in the Yelp data. Panels **(a)** and **(b)** show the distribution of the spreading concentration for the Yelp and model data, respectively.

6.3 Spreading concentration

Results for this metric on the Yelp data are shown in Figure 5(a). The values computed in the original network are easily distinguishable from the values computed in the randomized networks. It is also the case for the model data reported in Figure 5(b) where the curves progressively shift as p increases. Note that the items with the highest r_α values do not have particularly high degree. For instance, the average degree of the top 100 items in terms of r_α is 22 which is similar to the overall average degree which is 16. In other words, the friendship relations among the users who collect items of mediocre popularity are the most informative for assessing the strength of friends' influence in the data by the spreading concentration.

7 Results on the Digg data

We now apply the same metrics on data from the Digg web site to see how well our results translate to a different system. Figure 6 shows the results of applying the influence of friends metrics on the Digg data. The probability of collecting an item as a function of the normalized exposure shown in Figure 6(b) again better distinguishes the original data from their randomized counterparts than the raw exposure shown in Figure 6(a). Also the normalized contagion count shown in Figure 6(d) performs in this respect better than the raw contagion count shown in Figure 6(c). As shown in Figure 6(e), the difference between the original and randomized network is less pronounced when the spreading concentration metric is used. This is a consequence of the fact that our dataset only contains popular items that were promoted to the front page (the popularity of the least popular item is 111). We thus lack the items of average and little popularity that, as we discussed before, have the best chance to achieve high values of r_α by spreading within small groups of friends. The spreading concentration is thus not suitable for this kind of dataset.

8 Conclusion

Influence of friends is an important and omnipresent process and various metrics have been designed to quantify its strength in data. We use a data randomization technique to show that the existing metrics - exposure and the contagion count - perform poorly in the task for which they have been designed. We identified the broad distribution of item popularity and preferential attachment, common features in many real systems, as the main reason for this failure and we used it as a motivation for the design of new normalized metrics that are not biased by them. Our metrics are also and well motivated and supported

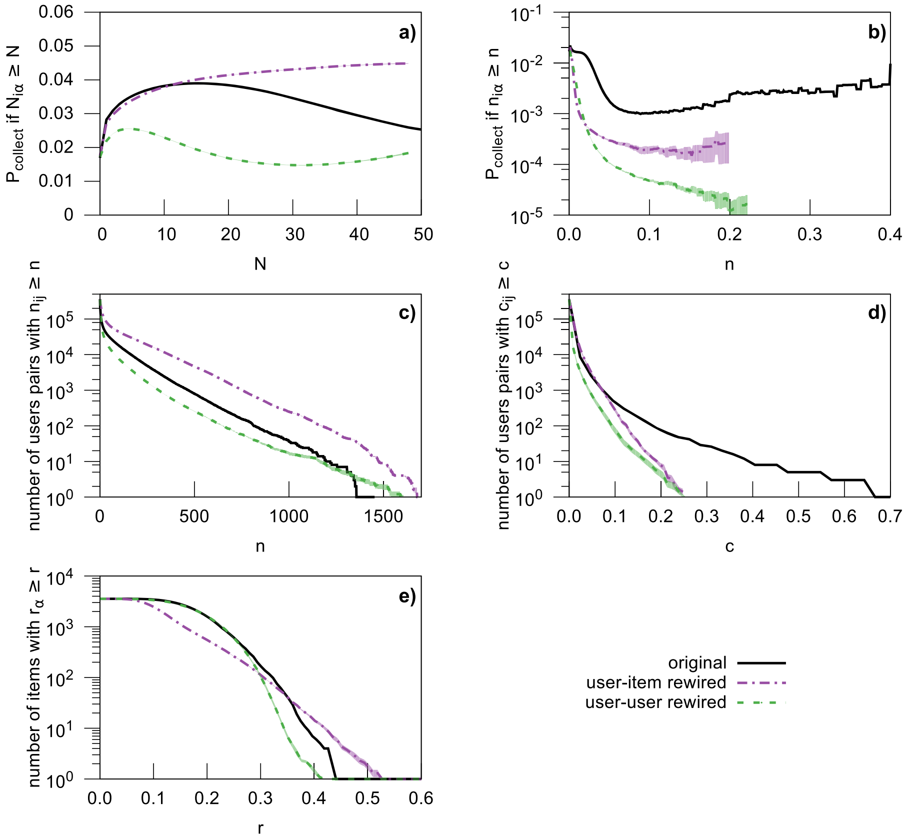

Figure 6 The metrics in the Digg data. Metrics of friends' influence applied on the Digg dataset and its randomized counterparts. Panels **(a)** and **(b)** show the item collection probability as a function of raw and normalized exposure, respectively. Panels **(c)** and **(d)** show the raw contagion and normalized count, respectively. Panel **(e)** shows the distribution of the spreading concentration among the items.

by the data. Our work opens new perspectives for the understanding of coupled networks and the design of future algorithms such as link prediction or community detection.

By rescaling the raw exposure of a user's friends by its activity, we introduce the normalized exposure and show that this metric is able to distinguish between the original and randomized data as well as to distinguish between model data generated with different values of the friends' influence parameter. Besides measuring influence of friends, this metric could be used to improve suggestions from web sites by assigning additional weight to items with a high value of normalized exposure or as an additional ingredient in recommender systems acting on both user-user and user-item networks. The metric could be further extended by taking into account the aging of exposure and thus assigning higher weights to recent activity of friends.

In a similar spirit, we find that the raw contagion count is biased by user activity and introduce a normalized contagion count by dividing with the expected contagion count in the null model. Since we find a high correlation between user activity and user degree in the social network, this normalization reduces the importance of influential highly connected users. Indeed, we find that the highest normalized contagion count is achieved by low or moderately active users who, due to the power-law distribution of user activity, are in majority in the data and their actions determine the evolution of the system. The newly proposed metric focuses on the behavior of these users and measures the impact of friends' influence on their behavior.

Similarly to moderately active users, small degree items account for a significant part of the data. Our third devised metric, spreading concentration, compares the user social network with the subnetwork consisting of users who collected a given item and reaches high values especially for niche items that spread locally among friends (i.e., the subnetworks corresponding to them are dense). While here we focus only on the overall distribution of spreading concentration, its values for individual items could be used, for example, to detect 'cult' items that have a special status inside specific communities. The time evolution of its value for individual items and its use in the identification of specific patterns early in an item's lifetime deserve further study in the future. A possible follow-up of the work could be to compare or combine our locally-defined measures to global ones, inspired by the recent studies on user influence [28] and multi-level networks [42].

We finally stress that our artificial model, although simple, reproduces several features observed in the real Yelp data. We used the model to show that the newly defined metrics are sensitive to the proportion of links driven by social interactions, but the original metrics fail in this respect. Given the level of agreement between the model and real data, the model itself can be used in the future to further our understanding and measurements of friends' influence as well as to study related aspects such as the presence and impact of influential users. In this work we have entirely neglected the time difference between the co-collection behavior of users. It is well possible that this information can be used to improve the measurement of friends' influence among the users.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AV, MM and YCZ designed the study. AV analyzed the data and prepared the figures. AV and MM wrote the main text of the manuscript. All authors approved the final version of the manuscript.

Acknowledgements

This work was supported by the Swiss National Science Foundation Grant No. 200020-143272 and by the EU FET-Open Grant No. 611272 (project Growthcom). We would like to thank Manuel Mariani for his careful proofreading of the manuscript.

Received: 26 August 2015 Accepted: 4 November 2015 Published online: 14 November 2015

References

1. Domingos P, Richardson M (2001) Mining the network value of customers. In: Proceedings of the 7th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 57-66
2. Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc R Soc Lond Ser A, Math Phys Sci 115:700-721
3. Leskovec J, Adamic LA, Huberman BA (2007) The dynamics of viral marketing. ACM Trans Web 1(1):5
4. Hodas NO, Lerman K (2014) The simple rules of social contagion. Sci Rep 4:4343
5. Shalizi CR, Thomas AC (2011) Homophily and contagion are generically confounded in observational social network studies. Sociol Methods Res 40(2):211-239
6. Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proc Natl Acad Sci USA 106(51):21544-21549
7. Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S (2008) Feedback effects between similarity and social influence in online communities. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 160-168
8. 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. ACM, New York, pp 7-15
9. Bass F (1969) A new product growth for model consumer durables. Manag Sci 15(1):215-227
10. Granovetter M (1978) Threshold models of collective behavior. Am J Sociol 83(6):1420-1443
11. Watts DJ (2002) A simple model of global cascades on random networks. Proc Natl Acad Sci USA 99(9):5766-5771
12. Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 137-146

13. Goldenberg J, Libai B, Muller E (2001) Talk of the network: a complex systems look at the underlying process of word-of-mouth. Mark Lett 12(3):211-223
14. Myers SA, Zhu C, Leskovec J (2012) Information diffusion and external influence in networks. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 33-41
15. Yang J, Leskovec J (2010) Modeling information diffusion in implicit networks. In: 2010 IEEE 10th international conference on data mining (ICDM), pp 599-608
16. Pálovics R, Benczúr AA (2013) Temporal influence over the Last.fm social network. In: 2013 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 486-493
17. Zheleva E, Sharara H, Getoor L (2009) Co-evolution of social and affiliation networks. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 1007-1016
18. Gomez Rodriguez M, Leskovec J, Schölkopf B (2013) Structure and dynamics of information pathways in online media. In: Proceedings of the 6th ACM international conference on web search and data mining. ACM, New York, pp 23-32
19. Newman ME (2001) Scientific collaboration networks. I. Network construction and fundamental results. Phys Rev E 64(1):016131
20. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L (2000) The large-scale organization of metabolic networks. Nature 407:651-654
21. Newman ME (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci USA 98(2):404-409
22. Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661-703
23. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509-512
24. Adamic LA, Huberman BA (2000) Power-law distribution of the world wide web. Science 287(5461):2115
25. Newman M (2009) The first-mover advantage in scientific publication. Europhys Lett 86(6):68001
26. Medo M, Cimini G, Gualdi S (2011) Temporal effects in the growth of networks. Phys Rev Lett 107(23):238701
27. Cha M, Haddadi H, Benevenuto F, Gummadi PK (2010) Measuring user influence in Twitter: the million follower fallacy. In: 4th international AAAI conference on weblogs and social media (ICWSM)
28. Morone F, Makse HA (2015) Influence maximization in complex networks through optimal percolation. Nature 524(7563):65-68
29. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. Nat Phys 6(11):888-893
30. Yelp Inc (2014) Yelp's academic dataset. http://www.yelp.com/academic_dataset
31. Hogg T, Lerman K (2012) Social dynamics of Digg. EPJ Data Sci 1:5
32. Romero DM, Meeder B, Kleinberg J (2011) Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In: Proceedings of the 20th international conference on world wide web. ACM, New York, pp 695-704
33. 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. ACM, New York, pp 721-730
34. Ver Steeg G, Ghosh R, Lerman K (2011) What stops social epidemics? In: 5th international AAAI conference on weblogs and social media (ICWSM)
35. Katona Z, Zubcsek PP, Sarvary M (2011) Network effects and personal influences: the diffusion of an online social network. J Mark Res 48(3):425-443
36. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 44-54
37. Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on Twitter. In: Proceedings of the 4th ACM international conference on web search and data mining. ACM, New York, pp 65-74
38. Bakshy E, Karrer B, Adamic LA (2009) Social influence and the diffusion of user-created content. In: Proceedings of the 10th ACM conference on electronic commerce. ACM, New York, pp 325-334
39. Goyal A, Bonchi F, Lakshmanan LV (2010) Learning influence probabilities in social networks. In: Proceedings of the 3rd ACM international conference on web search and data mining. ACM, New York, pp 241-250
40. Newman ME (2003) The structure and function of complex networks. SIAM Rev 45(2):167-256
41. Medo M (2014) Statistical validation of high-dimensional models of growing networks. Phys Rev E 89(3):032801
42. Radicchi F (2015) Percolation in real interdependent networks. Nat Phys 11(7):597-602