Information Fusion 28 (2016) 83–89

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Users joining multiple sites: Friendship and popularity variations across sites

Reza Zafarani [a,*], Huan Liu [b]

[a] Department of Electrical Engineering and Computer Science, Syracuse University, United States
[b] Computer Science and Engineering, Arizona State University, United States

ARTICLE INFO

Article history:
Received 18 November 2014
Received in revised form 25 May 2015
Accepted 4 July 2015
Available online 11 July 2015

Keywords:
Cross-site information fusion
Cross-site user analysis
Friendship analysis
Popularity analysis
Cross-media study

ABSTRACT

Our social media experience is no longer limited to a single site. We use different social media sites for different purposes and our information on each site is often partial. By collecting complementary information for the same individual across sites, one can better profile users. These profiles can help improve online services such as advertising or recommendation across sites. To combine complementary information across sites, it is critical to understand how information for the same individual varies across sites. In this study, we aim to understand how two fundamental properties of users vary across social media sites. First, we study how user friendship behavior varies across sites. Our findings show how friend distributions for individuals change as they join new sites. Next, we analyze how user popularity changes across sites as individuals join different sites. We evaluate our findings and demonstrate how our findings can be employed to predict how popular users are likely to be on new sites they join.

1. Introduction

Social media has become an integral part of our daily life. Its popularity has become evident with around 6 billion photos uploaded monthly to Facebook, the blogosphere doubling every five months, 72 h of video being uploaded every minute to YouTube, and Twitter having more than 200 million active users who tweet 500 million Tweets per day. According to a recent Pew Internet and American Life survey [1], more than 73% of online adults are on a social networking site. Clearly, our social media experience is no longer limited to a single site.

Our daily social media experience constitutes posting, liking, watching, listening, and the like on multiple sites such as Facebook, Twitter, Pandora, and YouTube. The same survey reports that a striking 42% of online users are now on multiple social media sites. This clearly shows the need for techniques that combine user information across sites. By combing the information that a user has provided across sites, one can better understand and profile the user, and in turn, improve online services such as recommendations to the user. However, it is not clear whether user information varies across sites. And if it does, how much does this information vary across sites. The answers to these questions are critical for a systematic user information fusion across sites. Our goal in this paper is to tackle this question.

As friends are the fundamental building blocks of social media sites [2], we focus on how friends and friendship behavior varies across sites. Friendship behavior and friends are naturally connected to the concept of popularity. Often, an intuitive mechanism to achieve popularity is to befriend others. Friends introduce a more pleasant social media experience and having more friends is perceived as a sign of popularity. For example, on social media, some individuals befriend random individuals in order to increase their popularity. Hence, we extend our study by analyzing both friendship behavior and popularity variations across sites. We show how friends are dispersed across sites and how this distribution shifts as users join more sites. We show how joining more sites influences the number of friends individuals have across them, as well as their popularity. Finally, we demonstrate how the findings of this study can be used to predict the popularity of users on new sites.

We first discuss the social media sites that users join. Next, we analyze how friends are distributed across sites. Then, we study how popularity varies across sites and detail our approach to predict user popularity across sites. Finally, we review related research to this study and conclude this work with future research directions.

2. Social media sites that users join

To understand user friendships and popularity across sites, one needs to gather the list of sites that users have joined on social

* Corresponding author. Tel.: +1 480 727 7349.
E-mail addresses: reza@zafarani.net (R. Zafarani), huan.liu@asu.edu (H. Liu).

http://dx.doi.org/10.1016/j.inffus.2015.07.002
1566-2535/

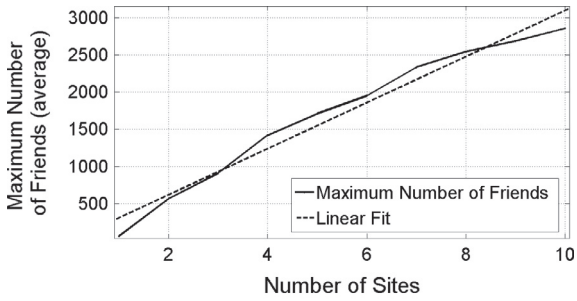
media. Social media sites are developed for different purposes; therefore, to systematically study friendships and popularity, one has to consider different types of sites. According to recent studies [3,4], social media sites can be categorized into 7 general categories: (1) *Blogs and Blogging Portals*, (2) *Media Sharing (Photo, Audio, or Video)*, (3) *Microblogging*, (4) *Social Bookmarking*, (5) *Social Friendship networks*, (6) *Social News and Search*, and (7) *Location-Based Networks*. We select 20 sites that cover these categories and are of different popularity on social media to study user friendships. The sites selected are *BlogCatalog*, *BrightKite*, *Del.icio.us*, *Digg*, *Flickr*, *iLike*, *IntenseDebate*, *Jaiku*, *Last.fm*, *LinkedIn*, *Mixx*, *MySpace*, *MyBlogLog*, *Pandora*, *Sphinn*, *StumbleUpon*, *Twitter*, *Yelp*, *YouTube*, and *Vimeo*. Next, we need to gather users that have joined some of these 20 sites.

Unfortunately, information about sites that users joined is not readily available. One can survey individuals and ask for the list of sites they have joined. This approach can be expensive and the data collected is often limited. Another method for identifying sites users have joined is to find users manually across sites. Users often provide personal information such as their real names, E-mail addresses, location, gender, profile photos, and age on different websites. This information can help find the same individual on different sites. However, finding users manually on sites can be challenging and time consuming. Automatic approaches are also possible that can connect corresponding users across different sites [5–11]. A more straightforward approach is to use websites where users voluntarily list the sites they have joined. In particular, we find social networking sites, blogging and blog advertisement portals, and forums to be credible sources for collecting the sites users have joined. For example, on social networking sites such as Google+ or Facebook, users can list their IDs on other sites. Similarly, on blogging portals and forums, users are often provided with a feature that allows users to list their usernames in other social media sites.

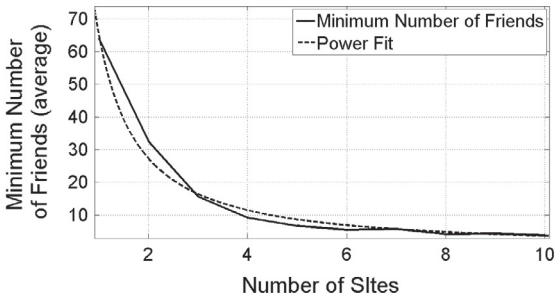
We utilize these sources for collecting sites users have joined. Overall, we collect a set of 96,194 users, each having accounts on some of the aforementioned 20 social media sites. As each user, has self-reported the accounts, they are guaranteed to belong to the same user. For each of the 20 sites, we develop a crawler that extracts the number of friends each individual has on the site. Hence, for each individual in our dataset, we have the number of friends a user has across different sites.

3. How friendship behavior varies across sites

One naturally expects that as users join more sites, it becomes more likely for them to find sites that contain more of their friends; therefore, befriending more individuals. Our data confirms this. Consider a user for whom we have his or her number of friends on n sites. Let $f_1, f_2, \ldots, f_n$ denote the number of friends of this user on these sites. Let $f_{\max} = \max(f_1, f_2, \ldots, f_n)$. We find $f_{\max}$ for all users in our dataset and group these users based on how many sites they have joined (n). We take the average $f_{\max}$ for users in each group. Fig. 1(a) plots the average maximum friend count $f_{\max}$ for users that have joined different numbers of sites (n). We observe that as users join more sites, their maximum friend count across sites on average increases. A linear line ($g(x) = 309.8x - 0.005177$), found with 95% confidence, fits to the curve with adjusted $R^2 = 0.9978$. R^2 is the *coefficient of determination* and $R^2 = 1$ denotes that a line perfectly fits the data. Hence, the expected maximum friend count across sites for users that have joined n sites is approximately n times more than that of users that have joined a single site. Similarly, one expects that as users join more sites, it becomes more likely for them to become inactive on some sites. Our data also confirms this. Fig. 1(b) shows



(a) Average Maximum Numbers of Friends.



(b) Average Minimum Numbers of Friends.

Fig. 1. Average minimum and maximum numbers of friends for users that have joined different numbers of sites.

the average minimum numbers of individuals befriended across sites as users join more sites. We observe a decrease in the minimum number of friends across sites as users join more sites. A power function ($g(x) = 65.03x^{-1.251}$), found with 95% confidence, fits this curve with adjusted $R^2 = 0.9878$. In other words, unlike the likelihood of having many friends that increases linearly as users join sites, the probability of having a few friends increases exponentially. Having said that, one can conjecture that (1) as the minimum friend count across sites is decreasing more sharply than the maximum, one should expect a decrease in the average number of friends individuals have across sites. As an alternative, one can conjecture that (2) the average number of friends should increase because the maximum number of friends individuals have across sites is much higher than the (few) number of friends they have on sites that they are inactive.

Our data shows that neither of these conjectures are valid for average numbers of friends across sites. Fig. 2 shows the average numbers of friends users have across sites as they join more sites. The figure shows that as users join more sites their average number of friends increases; however, once they join around 6 sites this average converges at around 400 friends. This average does not change much as users join new sites. This finding is in line with

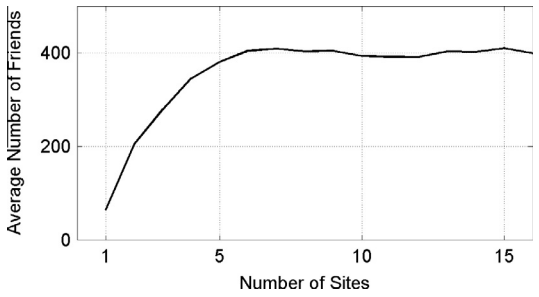


Fig. 2. Average numbers of friends for users that have joined different numbers of sites.

previous [12] and recent [13–15] literature on human cognitive limitations in maintaining communication and friendship with large groups of individuals.

There could be different explanations why the average of a distribution converges as we add more data points. For instance, by adding equally dispersed data points one can maintain the mean. To understand better how users befriend others, it is natural to observe how standardized moments of the friend count distribution changes. In particular, skewness [16], the third standardized moment $\left(\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]\right)$, and kurtosis [17,18], the forth standardized moment $\left(\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]\right)$, can help us understand why the average number of friends converges as users join more sites.

Skewness shows where the mass of the distribution is concentrated and whether the left or right tail of the distribution is longer. Skewness of 0 demonstrates a normal distribution where the mean is equal to the median. A positive skewness shows that while extreme values exist to the right of the distribution, the mass of the distribution is concentrated on the left of it. Negative skewness shows the opposite. For example, sample: {1,2,3,1000} has a positive skewness and sample: {1,1001,1002,1003} has a negative skewness. To account for small-sample bias, we compute the bias-corrected skewness for sample $x = (x_1, x_2, \ldots, x_n)$ as follows:

$$s = \frac{\sqrt{n(n-1)}}{n-2} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^3}, \tag{1}$$

where $\bar{x}$ is the mean for x. For each user, we compute the skewness of the user's friend counts across sites. Fig. 3 shows the empirical cumulative distribution function (Kaplan–Meier estimate) for these user skewness values for users that have joined different numbers of sites. We observe that most of the skewness values are positive showing that while there are extreme friend count values, the mass of the friend count distribution is concentrated on the left. Furthermore, we see that as users join more sites, the cumulative distribution function (CDF) moves to the right, showing that as users join more sites, the proportion of sites where they have fewer friends increases. In other words, users that have joined a few sites are more likely to be highly active on some sites compared to those users that joined more sites. Although we now know that users are more likely to have fewer friends on most sites they join, it is not known how these fewer friend counts are distributed. To observe where these fewer friend count values are concentrated, we measure the kurtosis of the distribution.

Kurtosis value of a distribution measures the peakedness of a probability distribution and how heavy-tailed it is. We use the bias-corrected kurtosis for small sample $x = (x_1, x_2, \ldots, x_n)$:

$$k = \frac{n-1}{(n-2)(n-3)}((n+1)k_0 - 3(n-1)) + 3, \tag{2}$$

where k_0 is

$$k_0 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}. \tag{3}$$

A kurtosis value of 3 shows a normal distribution and a value greater than 3 shows a *leptokurtic* distribution that has a more acute peak around the mean and more heavy tails. Similarly, a negative kurtosis value shows a *platykurtic* distribution with a less pronounced and wider peak. For each user, we compute the kurtosis of the user's friend counts across sites. Fig. 4 shows the empirical cumulative distribution (Kaplan–Meier estimate) for these user kurtosis values for users that have joined different numbers of sites. The graph shows that most kurtosis values are more than 3, denoting that the users' friend counts are more concentrated around the mean than normally expected. Furthermore, we observe that the CDF curves move to the right for users that have joined more sites. In other words, users' friend counts across sites tend to concentrate more around the mean value as users join more sites. Since we know from skewness analysis that users befriend a few others on most sites they join, this shows that the number of few individuals befriended are concentrated around a mean value. In other words, each user has almost the same number of friends (e.g., 10 friends) across most sites. The mean value varies for different users.

The initial increase in the average number of friends shows that when users join a few sites, it is more likely for them to get engaged while befriending many; however, as they join more sites, they start to become inactive in those sites and the average converges.

4. How popularity changes across sites

We have analyzed how the number of friends varies across sites. In this section, we perform similar experiments to analyze how user popularity changes across sites. To measure popularity we note that users with many friends are often considered popular users. So, a natural way to quantify popularity on a site is to use individual's friend count. However, the same number of friends on different social networks implies different levels of popularity due to different distributions of friend counts. For comparison, one can simply convert the friend count to the probability of observing the friend count, which is comparable across sites. A lower probability indicates a higher popularity. It is well known that the distribution of friend counts in a social media site often follows a power-law distribution [19,20]. Hence, we perform the

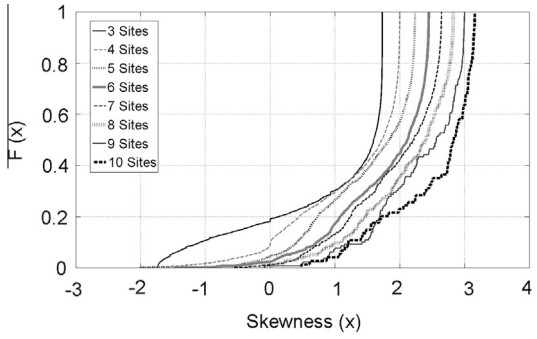


Fig. 3. Empirical cumulative distribution for skewness of friend distribution as users join more sites.

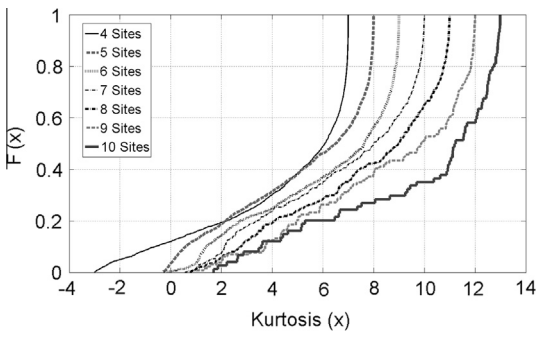


Fig. 4. Empirical cumulative distribution for kurtosis of friend distribution as users join more sites.

systematic procedure outlined in [21] for each of our 20 sites to determine their parameters for the power-law distribution. For integer values, the power-law distribution is defined as

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})}, \tag{4}$$

where

$$\zeta(\alpha, x_{min}) = \sum_{n=0}^{\infty} (n + x_{min})^{-\alpha} \tag{5}$$

is the generalized Hurwitz zeta function, α is the power-law exponent and x_{min} is the minimum value for which for all $x \geqslant x_{min}$, the power-law distribution holds. We estimate α and x_{min} using a finite sample correction bias using the maximum likelihood method outlined in [21]. Given these parameters, for any friend count $f \geqslant x_{min}$, we estimate the probability of observing f (i.e., $p(x = f)$) using Eq. (4).

Recent studies show that using the power-law distribution may not be always appropriate for modeling the friend count distribution of social networks [22,23]. Hence, when $f < x_{min}$, instead of using Eq. (4), we use the maximum likelihood estimate of $p(f)$,

$$p(f) = \frac{n_f}{n}, \tag{6}$$

where n_f is the number of users on the site with f friends and n is the total number of users on the site.

Following this approach, we estimate the probability of observing all friend counts in our dataset; hence, having the popularity of all users in our data across sites. Given these user popularity values across sites, we first measure how the average popularity varies across sites. Fig. 5 provides average popularity for users that have joined different numbers of sites. Notice that convergence also takes place for user popularities. Users are least popular when they have joined a single site and they are most popular, when they have two or more accounts. Popularity saturates much faster and as users join sites, their average popularity remains unchanged.

While the average popularity shows how users popularity changes across sites on average, it does not show how a user's popularity changes as he or she joins new sites. This is because we have no temporal information on what sites were joined first and how popularity increased or decreased over time. However, one can approach this problem by computing the expected popularity change over time.

Consider a user for whom we have his or her number of friends on n sites. Let $f_1, f_2, \ldots, f_n$ denote the number of friends of this user on these sites. Among the n sites that the user has joined, there must be a site that is joined after all others. Since we have no temporal information, the last site could be any of the n sites. We consider n cases. In each case, we consider one of the sites as the last site that the user has joined and the other $n-1$ sites as the sites that the user has joined in the past. In case $1 \leqslant i \leqslant n$, we consider

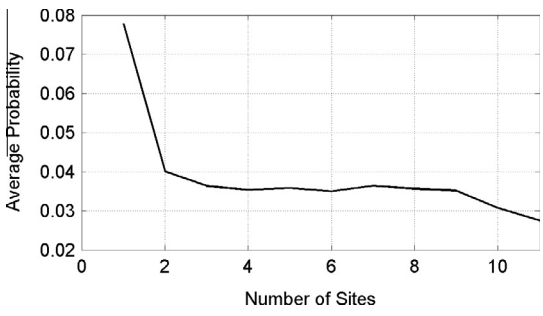


Fig. 5. Average popularity for users that have joined different numbers of sites.

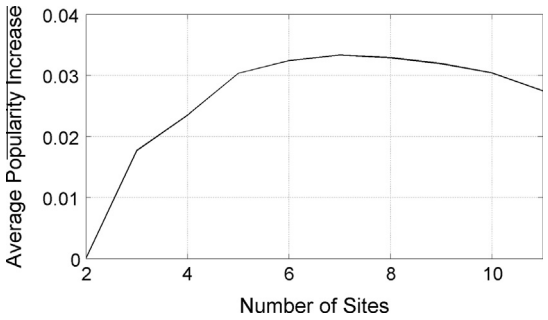


Fig. 6. Average popularity increase for users that have joined different numbers of sites.

that the user in the $n-1$ sites has $f^i_1, f^i_2, \ldots, f^i_{n-1}$ friends and f^i_n friends on the last site. The popularity values can be estimated by computing the probability of observing each friend count: $p(f^i_1), p(f^i_2), \ldots, p(f^i_n)$. For the $n-1$ sites that the user has joined, the maximum popularity that the user achieved is $\min(p(f^i_1), p(f^i_2), \ldots, p(f^i_n))$. The user has become more popular on the nth site if and only if,

$$\min(p(f^i_1), p(f^i_2), \ldots, p(f^i_n)) < p(f^i_n). \tag{7}$$

Thus, we measure popularity increase for case i as

$$p(f^i_n) - \min(p(f^i_1), p(f^i_2), \ldots, p(f^i_n)). \tag{8}$$

Since, the last site that a user joined is not known, we compute the expected *popularity increase* as

$$\frac{1}{n}\sum_{i=1}^{n}\left[p(f^i_n) - \min(p(f^i_1), p(f^i_2), \ldots, p(f^i_n))\right]. \tag{9}$$

The average expected popularity increase for users that have joined different numbers of sites is provided in Fig. 6. The figure shows that users tend to increase their popularity faster as they join more sites; however, there is a cap to the level at most a user can increase his or her popularity and this level is as users join 7 sites.

5. Predicting user popularity

We have demonstrated that user friendships and popularity exhibits specific patterns as users join sites. This brings about a challenging, yet unexplored question: can one predict user's popularity on a new site?

Predicting user's popularity can not only help recommend new sites to users as they search for new sites on the web, but more importantly, can help sites identify users that are more likely to be interested in joining and becoming active on them. One expects a rather complicated solution to this problem. An approach that has access to different types of information and users' interests and a matching procedure that identifies sites on which users are most likely to become active. Even then, one needs to know if the site includes friends of an individual for better popularity prediction.

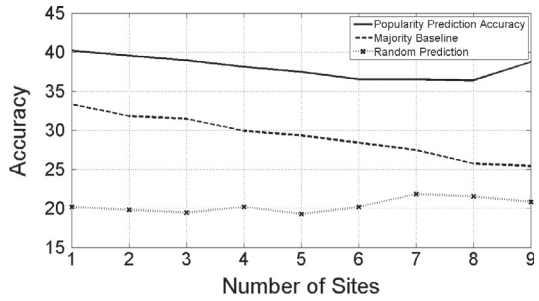
If the popularity patterns is our data were meaningless, one should not be able to observe their effect in predicting user's popularity. By extracting popularity patterns a user has exhibited in the past, one can predict the popularity of a user in the future. In this section, we demonstrate how one can use **only** popularity patterns and outperform baseline methods that use no popularity patterns, safely concluding that the obtained popularity patterns can be used to predict user's popularity.

For any user in our dataset that has joined n sites, we assume that given the user's popularity level on $n-1$ of these sites, the popularity of the nth site should be predictable. To determine the popularity level of users in sites, we divide the users on each site into five categories. These categories are based on the level of popularity and their proportion are inspired by the diffusion of innovations theory [24], where individuals depending on their time of adopting a new product are categorized into 5 categories: innovators (top 2.5%), early adopters (next 13.5%), early majority (next 34%), late majority (next 34%), and laggards (last 16%). For each site, we divide users into 5 categories based on their level of popularity: elites, highly popular, averagely popular, averagely unpopular, and unpopular users. We use popularity categories instead of the actual probability as this introduces a generalizable prediction algorithm as users with different probabilities and new sites appear on social media. Thus, for each user that has joined n sites, we generate all the $\binom{n}{n-1} = n$ combinations of $n-1$ sites as historical data. For each combination, we construct a data instance of 5 features, each representing a popularity level. For each popularity level, we count the number of sites the user has joined in the past among his or her $n-1$ sites and has expressed that level of popularity. We set the class label as the popularity level for the user in the nth site (i.e., a value in {1,2,3,4,5}). We generate 39,130 instances. Our initial attempt to predict the class label in this dataset using Naive Bayes classifier predicts user popularity with an accuracy of 38.50% and an AUC of 0.618. The area under the ROC curve (AUC) is a criterion used to measure the quality of a classification algorithm and ranges between 0.5 and 1. To determine the sensitivity of our results to the learning bias of different algorithms, we test a variety of classification techniques. The results are provided in Table 1. We observe minimal sensitivity to learning bias, showing that one can reasonably predict user's popularity regardless of the classification algorithm. Logistic Regression performs the best with 39.26% accuracy in predicting user popularity and an AUC of 0.627. Thus, logistic regression is used for the rest of our experiments. Note that in the dataset constructed, the probabilities were converted to popularity categories (i.e., discretized) based on unequal frequencies. Hence, we also experimented when binning probabilities with equal frequencies (20% intervals). While the performance dropped in this new dataset for logistic regression with an accuracy of 27.29% and an AUC of 0.605, the difference was not significant. Therefore, we continued our experiments with the initial dataset were probabilities were discretized according to diffusion of innovations theory.

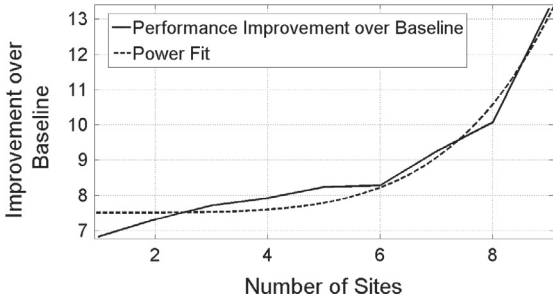
In our data, users have joined different numbers of sites. To verify helpfulness of adding more sites on user popularity prediction, we partition our dataset. Partition i contains the set of users that have already joined i sites. We perform popularity prediction for each partition. Fig. 7(a) shows that the prediction results (accuracy) for each partition does not variate much. The figure also shows as a dashed line the majority class predictor for each partition and the random prediction results. The majority baseline predicts the popularity level of **all** users in the partition as the popularity level that is most common among the (training) users of that partition. Since the partitions were slightly imbalanced, we also computed the AUC and found that it was mostly fixed with an average AUC of 0.6273. The same figure shows that for all cases, we outperform the majority predictor, proving that popularity patterns across sites can help predict the popularity of a user on a new site.

Table 1
Site recommendation performance.

Technique	AUC	Accuracy (%)
Logistic regression	0.627	**39.26**
SMO (sequential minimal optimization)	0.574	38.84
J48 decision tree learning	0.604	38.82
Random forest	0.612	38.63
Naive Bayes	0.618	38.50



(a) Popularity Prediction Accuracy as Users Join Different Numbers of Sites.



(b) Performance Improvement over Baseline.

Fig. 7. Performance for popularity prediction.

Fig. 7(a) also shows that as users join more sites and more information becomes available the gap between the prediction outcome and the majority class starts to increase. The gap increase is provided in Fig. 7(b). The gap increases exponentially, fitting a power function ($g(x) = 8.65 \times 10^{-5}x^{5.068} + 7.506$) with adjusted $R^2 = 0.9494$. In other words, as more popularity patterns of a user becomes available to the prediction algorithm, one can predict user's popularity exponentially better.

6. Related work

Studying friendships and popularity on social media sites has a long history. The friendship network and popularity is often studied on a single site. Other related areas to this work are (1) analyzing dynamics of multiple networks and (2) analyzing user behavior across social media. We briefly review related research from each of these three areas and outline how this work stands compared to its related work.

6.1. Single-site friendship and popularity analysis

When considering only the number of friends individuals have, the analysis boils down to analyzing the degree distribution of social networks [25,26]. It has been shown multiple times that the degree distribution of these social networks follows a power-law distribution [27,28]. This study follows a similar approach; however, at a multi-site level, where we analyze how number of friends (degrees) changes across sites. Unlike the common degree distribution analysis where millions of nodes are analyzed to determine the degree distribution, with multiple sites, the number of available samples is limited to a few numbers. Hence, we take a different approach in this paper by observing how the

number of friends change across sites with the help of statistical measures.

6.2. Analyzing dynamics of multiple networks

Comparing network characteristics of multiple networks has been the subject of recent studies [29–31]. For instance, Mislove et al. [29] analyze 4 networks: Flickr, YouTube, LiveJournal, and Orkut and demonstrate that these networks exhibit various properties such as being scale-free and having a densely connected core of high-degree nodes. Although these studies analyze multiple networks, the analysis is performed irrespective of the users that are shared across networks. Our work focuses on how friends of shared users across networks are distributed and how popularity for the users changes across social media sites.

6.3. Analyzing user behavior across sites

Considering befriending as a behavior of individuals, the recent studies that analyze user behavior across sites becomes relevant to this work. Some studies analyze how a specific behavior changes across sites without considering users that are shared across sites [32]. Other recent studies consider a specific behavior across sites such as Tagging [33,34], but for users that are shared across sites. Our work is related to both as it analyzes the variation of an unexplored behavior (i.e., befriending) and user popularity across sites for users shared across sites.

7. Conclusions and future work

Social media users are members of multiple sites. For a systematic study of users on social media one has to combine their information across sites. In this study we investigate how this information varies across sites. We focus on the most fundamental information available across social media sites: user friends and their popularity.

By studying user friendships and popularity across sites, we showed that the maximum number of friends individuals have across sites increases linearly as users join sites and their minimum drops exponentially. Furthermore, we noticed that as users join sites their average number of friends converges to a value near 400. We investigated this phenomenon even further and showed that as users join sites, the likelihood of observing fewer friend counts increases and at the same time, users frequently exhibit their mean behavior, such as always befriending 10 people. This frequent behavior of befriending a few friends on most sites leads to users converging to an average of 400 friends across sites.

By computing the power-law distribution parameters for these sites, we computed user popularity on sites. We found that popularity follows the same trend as in friend counts, converging to an average value. This result shows that users joining multiple sites cannot increase their average popularity and that the average popularity converges to a fixed value as users join sites. We also demonstrated that as users join sites, the amount their popularity can increase has a constant upper bound. Finally, we showed how the popularity patterns of users can be used to determine their popularity on future sites. Using a straight-forward approach we showed that as patterns of popularity become available to the popularity prediction algorithm, the algorithm gains exponential performance gain over baselines.

The study presented in this paper has multiple implications. It not only allows for popularity prediction across sites, but can also be combined with studies that predict what sites users will join [35]. Once sites that users will join are predicted, methods discussed in this paper can help identify users that are more likely to become popular, which in turn can help sites determine users with the highest priority for friend recommendation algorithms [2].

While data collection for our study was challenging, we believe with more data regarding the behavior and interests of users across sites, one should be able to obtain deeper insights into how users change behavior across sites and improve the performance of site popularity prediction. While our data did not contain temporal information, with temporal information, one can effectively measure how joining a site influences popularities on other sites and how increase in one site's popularity can influence popularities on other sites. Furthermore, one can cluster sites based on popularity patterns and predict popularity in future sites based on the category of sites the new site belongs to. We consider these as promising future directions for this work.

References

[1] M. Duggan, A. Smith, Social media update 2013, Pew Internet Am. Life Project (2013).
[2] R. Zafarani, M.A. Abbasi, H. Liu, Social Media Mining: An Introduction, Cambridge University Press, 2014.
[3] A.M. Kaplan, M. Haenlein, Users of the world, unite! the challenges and opportunities of social media, Business Horizons 53 (1) (2010) 59–68.
[4] N. Agarwal, Social computing in blogosphere, Ph.D. thesis, Arizona State University, 2009.
[5] T. Iofciu, P. Fankhauser, F. Abel, K. Bischoff, Identifying users across social tagging systems, in: ICWSM, 2011.
[6] D. Perito, C. Castelluccia, M.A. Kaafar, P. Manils, How unique and traceable are usernames?, in: Privacy Enhancing Technologies, Springer, 2011, pp 1–17.
[7] R. Zafarani, H. Liu, Connecting users across social media sites: a behavioral-modeling approach, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2013, pp. 41–49.
[8] R. Zafarani, H. Liu, Connecting corresponding identities across communities, in: ICWSM, 2009.
[9] A. Malhotra, L. Totti, W. Meira Jr., P. Kumaraguru, V. Almeida, Studying user footprints in different online social networks, in: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, 2012, pp. 1065–1070.
[10] N. Korula, S. Lattanzi, An efficient reconciliation algorithm for social networks. Available from: <arXiv:1307.1690>.
[11] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, H.-W. Hon, What's in a name?: an unsupervised approach to link users across communities, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, ACM, 2013, pp. 495–504.
[12] R.I. Dunbar, Neocortex size as a constraint on group size in primates, J. Hum. Evol. 22 (6) (1992) 469–493.
[13] A. Kluth, Primates on facebook, The Econ. (2009).
[14] B. Gonçalves, N. Perra, A. Vespignani, Modeling users' activity on twitter networks: validation of dunbar's number, PloS One 6 (8) (2011) e22656.
[15] G. Miritello, R. Lara, M. Cebrian, E. Moro, Limited communication capacity unveils strategies for human interaction, Sci. Rep. 3 (2013).
[16] S.R.A. Fisher, S. Genetiker, R.A. Fisher, S. Genetician, G. Britain, R.A. Fisher, S. Généticien, Statistical Methods for Research Workers, vol. 14, Oliver and Boyd Edinburgh, 1970.
[17] L.T. DeCarlo, On the meaning and use of kurtosis, Psychol. Methods 2 (3) (1997) 292.
[18] H. Cremér, Mathematical Methods of Statistics (PMS-9), vol. 9, Princeton University Press, 1999.
[19] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[20] M. Newman, Networks: An Introduction, Oxford University Press, 2009.
[21] A. Clauset, C.R. Shalizi, M.E. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661–703.
[22] M. Gjoka, M. Kurant, C.T. Butts, A. Markopoulou, Walking in facebook: a case study of unbiased sampling of OSNs, in: Proceedings of the 29th Conference on Information Communications, IEEE Press, 2010, pp. 2498–2506.
[23] J. Ugander, B. Karrer, L. Backstrom, C. Marlow, The anatomy of the facebook social graph. Available from: <arXiv:1111.4503>.
[24] E.M. Rogers, Diffusion of Innovations, Simon and Schuster, 2010.
[25] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web, Comput. Netw. 33 (1) (2000) 309–320.
[26] J.M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A.S. Tomkins, The web as a graph: measurements, models, and methods, in: Computing and Combinatorics, Springer, 1999, pp. 1–17.
[27] D. Easley, J. Kleinberg, Networks, Crowds, and Markets, vol. 8, Cambridge Univ. Press, 2010.
[28] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology, ACM SIGCOMM Computer Communication Review, vol. 29, ACM, 1999, pp. 251–262.

[29] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, B. Bhattacharjee, Measurement and analysis of online social networks, in: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, ACM, 2007, pp. 29–42.
[30] J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins, Microscopic evolution of social networks, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2008, pp. 462–470.
[31] P. Shakarian, D. Paulo, Large social networks can be targeted for viral marketing with small seed sets, in: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, 2012, pp. 1–8.
[32] F. Benevenuto, T. Rodrigues, M. Cha, V. Almeida, Characterizing user behavior in online social networks, in: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, ACM, 2009, pp. 49–62.
[33] P.d. Meo, E. Ferrara, F. Abel, L. Aroyo, G.-J. Houben, Analyzing user behavior across social sharing environments, ACM Trans. Intell. Syst. Technol. (TIST) 5 (1) (2013) 14.
[34] F. Abel, E. Herder, G.-J. Houben, N. Henze, D. Krause, Cross-system user modeling and personalization on the social web, User Model. User-Adapted Interact. 23 (2–3) (2013) 169–209.
[35] R. Zafarani, H. Liu, Users joining multiple sites: distributions and patterns, in: Eighth International AAAI Conference on Weblogs and Social Media, 2014.