Users Joining Multiple Sites: Distributions and Patterns

Reza Zafarani and Huan Liu
Computer Science and Engineering
Arizona State University
{Reza, Huan.Liu}@asu.edu

Abstract

The rise of social media has led to an explosion in the number of possible sites users can join. However, this same profusion of social media sites has made it nearly impossible for users to actively engage in all of them simultaneously. Accordingly, users must make choices about which sites to use or to neglect. In this paper, we study users that have joined multiple sites. We study how individuals are distributed across sites, the way they select sites to join, and behavioral patterns they exhibit while selecting sites. Our study demonstrates that while users have a tendency to join the most popular or trendiest sites, this does not fully explain users' selections. We demonstrate that peer pressure also influences the decisions users make about joining emerging sites.

Our life in social media is no longer limited to a single site. We post on Reddit, like on Facebook, tweet on Twitter, watch on YouTube, listen on Pandora, along with many other activities exhibited by social media users. With the constant rise of new sites and advancement of communication technology, thousands of social media sites are at our fingertips. With so many choices, our attention spans are decreasing rapidly. On average, a user spends less than a minute on an average site (BBC News 2002). With our limited time and short attention span, we often face a dilemma of choosing a handful of sites over others. How do we select these sites?

As social media consumers, we are constantly seeking sites that can keep our attentions glued to our screens by providing engaging content, especially content generated by our friends. It is well-known that the likelihood of engaging in an activity is increased as more friends become engaged in that activity (Backstrom et al. 2006). Thus, it is natural to assume that users select sites where they find more friends on. On average, sites with more members are expected to contain more friends for an average individual; hence, it is expected for the users' site selection to be statistically biased toward more popular sites.

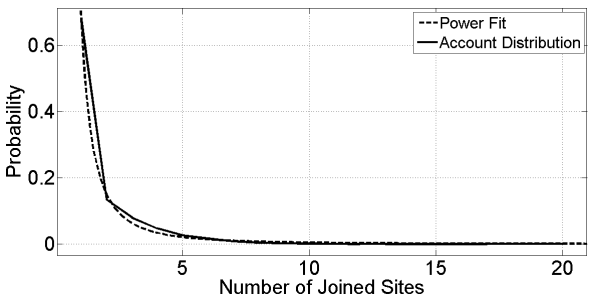
In this paper, we analyze users joining multiple sites. We show how users are dispersed across sites. By studying users across sites, we show that while there is a tendency to join popular sites, users exhibit a variety of site selection patterns. Finally, we evaluate the obtained users' site selection patterns with an application that recommends new sites to users for joining. Our evaluation demonstrates promising results and reveals additional interesting user joining patterns.

We first detail the data collection for our research. Next, we analyze user distribution across sites. Then, we outline membership patterns across sites, followed by our evaluation of these patterns. Finally, we conclude this work with a brief literature review and future directions.

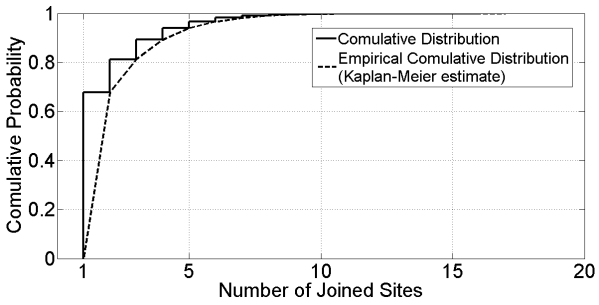
Data Preparation

To study user memberships across sites, one needs to gather sites that users have joined on social media. Unfortunately, this information is not readily available. One can simply 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 that users have joined is to find users manually across sites. Users, more often than not provide personal information such as their real names, E-mail addresses, location, gender, profile photos, and age on these websites. This information can be employed to 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 using minimum information such as their usernames (Zafarani and Liu 2013). A more straightforward approach is to use websites where users have the opportunity to list the sites they have joined. In particular, we find social networking sites, blogging and blog advertisement portals, and forums to be valuable sources for collecting the sites users have joined. For example, on most 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 utilized these sources for collecting sites that users have joined. Overall, we collected a set of 96,194 users, each having accounts on a subset of 20 social media sites. The sites included in our dataset 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*. The data was collected in 2008. In 2008, MySpace was the most important social networking site, BlogCata-



(a) Probability Distribution



(b) Cumulative Probability Distribution and Empirical Cumulative Distribution

Figure 1: Distribution of Users across Sites

log was one of the most popular blogging sites with social networking capabilities, and LinkedIn and Yelp were quite unpopular. At the time, Yelp had only 3 million users and LinkedIn was an order of magnitude smaller.

User Membership Distribution across Sites

First, we determine how users are distributed across sites. A natural way to determine the user distribution is to compute the proportion of users that have joined different number of sites. Figure 1(a) shows how users are distributed with respect to the number of sites they have joined. Figure 1(b) plots the cumulative distribution function and the empirical cumulative distribution function (Kaplan-Meier estimate) for the distribution in Figure 1(a). These figures show that more than 97% of users have joined at most 5 sites and users exist on as many as 16 sites.

A power function, $g(x) = 0.6761x^{-2.157}$, found with 95% confidence, fits to the distribution curve in Figure 1(a) with adjusted $R^2 = 0.9978$. The exponent -2.157 denotes that individuals that are members of n sites are $1/n^{2.157}$ less likely than individuals that are members of only one site. For example, users that are members of $n = 7$ sites are $\approx 1/66$ times less likely than users that are members of only one site. The power function fit is highly correlated to our data, indicating the possibility of a power-law distribution. To investigate this possibility, we follow the systematic procedure outlined in (Clauset, Shalizi, and Newman 2009) to determine whether the user distribution across sites follows a power-law distribution. For integer values, the power-law distribution is defined as

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})}, \tag{1}$$

where, $\zeta(\alpha, x_{min}) = \sum_{n=0}^{\infty}(n + x_{min})^{-\alpha}$ is the generalized Hurwitz zeta function, α is the power-law exponent and x_{min} is the minimum value for which for all $x \geq x_{min}$, the power-law distribution holds. We estimate α and x_{min} using the maximum likelihood method outlined in (Clauset, Shalizi, and Newman 2009). Our results shows that the value of α is slightly larger than the initially obtained exponent of 2.157 and is around 2.34. To verify the validity of our power-law fit, we calculate p-value using the Kolmogorov-Smirnov goodness-of-fit test. We obtain $p \approx 0$, rejecting the null hypothesis, showing that users across sites are distributed according to a power-law distribution.

User Membership Patterns across Sites

We showed that user distribution across sites is power-law. However, it is still unknown how users select sites to join. A common perception is that users are more likely to join most popular sites. Here, we show that this is not true in general. While there is a tendency to join popular sites, users exhibit different site selection patterns on social media.

Assume that sites are represented using a *complete* weighted graph $G(V, E, O)$. In this graph, nodes $v \in V$ represent sites. Let $|V| = n$. In our data, $n = 20$. An edge exists between all pairs of nodes, i.e., $E = V \times V$. Edge $e_{ij} \in E$ between two sites (nodes) i and j has weight $O_{ij} \in O$, where $O \in \mathbb{R}^{n \times n}$. Weight O_{ij} denotes the number of users that are members of both sites i and j. Let $O_{ii} = 0$.

Our collected dataset can be represented using a matrix $U \in \mathbb{R}^{l \times n}$, where l is the number of users. $U_{ij} = 1$, when user i is a member of site j and $U_{ij} = 0$, otherwise. Clearly, O matrix can be written in terms of U matrix,

$$O = (J_n - I_n) \circ U^T U, \tag{2}$$

where $J_n \in \mathbb{R}^{n \times n}$ is the matrix of all ones, I_n is the identity of size n, and $\circ$ is the Hadamard (entrywise) product.

For site v, let d_v represent the number of users that are on site v.[1] We can estimate[2] d_v as $d_v \approx \sum_i O_{vi}$.

For two sites i and j, we compute the number of users that are expected to be members of both. Assume that users randomly join a site with a probability that is proportional to its popularity. For any user in site i, the probability that the user joins site j is $\frac{d_j}{\sum_k d_k} = \frac{d_j}{2m}$, where $m = \frac{1}{2}\sum_k d_k$. As site i has d_i users, the *expected* number of members of both sites is $\frac{d_i d_j}{2m}$. The actual number of members of both sites is given in our data as O_{ij}. The distance between this actual number and its expected value ($O_{ij} - \frac{d_i d_j}{2m}$) indicates how non-random joining both i and j is. We expect the users' site selection behavior to be non-random. Thus, we can find communities of sites such that this distance is maximized for the

[1] This is equivalent to a node's degree in an unweighted graph.

[2] The estimation performs well in our setting and is close to the actual d_v; however, it considers independence among site overlaps.

sites in each community. These communities represent sites that users often join together. Let $P = (P_1, P_2, \ldots, P_k)$ denote a partitioning of the sites in V into k partitions. For partition P_x, this distance can be defined as

$$\sum_{i,j \in P_x} (O_{ij} - \frac{d_i d_j}{2m}). \quad (3)$$

This distance can be generalized for the partitioning P,

$$\sum_{x=1}^{k} \sum_{i,j \in P_x} (O_{ij} - \frac{d_i d_j}{2m}). \quad (4)$$

This summation term takes a maximum value of $\sum_{ij} O_{ij} \approx \sum_k d_k = 2m$; therefore, the normalized version of this distance is defined as

$$Q = \frac{1}{2m} [\ \sum_{x=1}^{k} \sum_{i,j \in P_x} (O_{ij} - \frac{d_i d_j}{2m})\ ]. \quad (5)$$

This is in fact a weighted version of the modularity measure defined by Newman (Newman 2006). We define the modularity matrix as $B = O - \mathbf{d}\mathbf{d}^T/2m$, where $\mathbf{d} \in \mathbb{R}^{n \times 1}$ is a vector that contains the number of members for all sites. Then, weighted modularity can be reformulated as

$$Q = \frac{1}{2m} \mathrm{Tr}(X^T B X), \quad (6)$$

where $X \in \mathbb{R}^{n \times k}$ is the partition membership matrix, i.e., $X_{ij} = 1$ iff. $v_i \in P_j$. This objective can be maximized such that the best membership function is obtained with respect to weighted modularity. Unfortunately, the problem is NP-Hard. Relaxing X to $\hat{X}$ that has an orthogonal structure ($\hat{X}^T \hat{X} = I_k$), the optimal $\hat{X}$ can be computed using the top k eigenvectors of B corresponding to positive eigenvalues.

Even when maximizing weighted modularity on our data, we obtain a negative value. The negative modularity denotes that users on average have other preferences when joining new sites than just selecting random popular sites.

Figure 2 shows the categorization of sites obtained using weighted modularity maximization. We observe several patterns in this figure. First, we notice that there are popular sites that users become members of all (or most). These sites are shown on the top right part of the figure in light orange. This cluster is MySpace, BlogCatalog, Twitter, and YouTube. For instance, we become members of Facebook to socialize with our friends, Twitter to post microblogging messages, YouTube to watch videos, and WordPress to write blogs. Back in 2008, MySpace and BlogCatalog were exemplars of prominent social networking and blogging sites. We believe this cluster of sites represent the average behavior of most users that are members of a few sites to satisfy their basic needs. The second group of sites are shown in the bottom part of the figure using green and red nodes. Green nodes represent audio/video/photo sharing sites such as online radios or video sharing sites that consumers often join **all** to be able to access the content that becomes available on each one of them. Similarly, the red nodes represent social tagging/social news/content sharing sites where individuals visit **all** to obtain interesting content. Reddit is a current popular example of these sites. The final group of sites shown in Blue, are unknown or unpopular sites that users rarely join. These are sites that are often joined by early adopters who wish to explore more and find new content or sites.

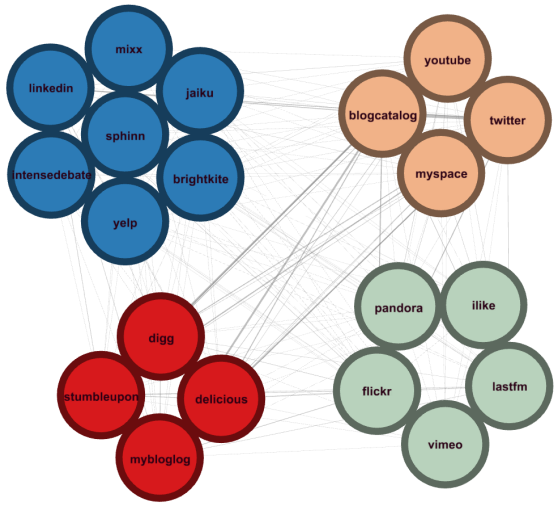


Figure 2: Site Categorization based on Sites that are Commonly Joined by Users.

Note that Yelp and LinkedIn were members of this cluster in 2008, which is due to their less popularity at that time. Note that these patterns are based on sites that are joined together; therefore, they are not mutually exclusive. A user can join sites in one or all of these clusters. Furthermore, a user should not necessarily be a member of all sites in each cluster, but can be a member of a subset of the sites.

After user membership patterns are obtained, it is imperative to validate these patterns. Because ground truth of the patterns is unavailable, one way of evaluating is to check if the patterns can help in some applications such as prediction or recommendation. In the following, we adopt the recommendation task as an evaluation strategy. As we will see, this approach leads to the further discovery of interesting patterns on how users select sites to join.

Evaluating via Recommending Sites to Users

If site selection patterns are not true patterns (i.e., random patterns), one should not be able to observe their effect in recommending sites to users. By identifying the types of site selection patterns a user has exhibited in the past, one can recommend sites to the user in the future. By outperforming baseline methods that use no user patterns, once can safely conclude that the obtained patterns are true patterns.

For any user in our dataset that has joined n sites, we assume that given the category (node color in Figure 2) of $n-1$ of these sites, the category of the nth site should be predictable. We use categories instead of the sites as this introduces a generalizable recommendation algorithm as 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 of $n-1$ sites, we construct a data instance of 4 features by counting the number of sites in each category that the user has joined in the past. This instance describes the amount of interest the user has expressed in each category in the past. We set the class label as the category of the nth site (i.e., a value

Table 1: Site Recommendation Performance

Technique	AUC	Accuracy
J48 Decision Tree Learning	0.880	**79.25%**
Random Forest	0.895	79.17%
Logistic Regression	0.886	79.14%
SMO (Sequential Minimal Optimization)	0.728	78.92%
Naive Bayes	0.869	76.66%

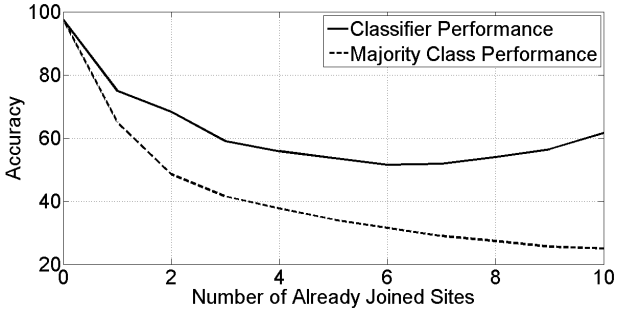


Figure 3: Recommendation Performance when the User has Already Joined some Sites.

in {1,2,3,4}). We generate 73,001 instances. Our initial attempt to predict the class label in this dataset using Naive Bayes classifier recommends a new site with an accuracy of 76.66% and an AUC of 0.869. 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 the learning bias. J48 performs the best with 79.25% accuracy in predicting the correct site category and an AUC of 0.88. Thus, J48 is used for the rest of our experiments.

To verify the influence of historical data on our results, we select 11 subsets of our dataset. Subset i, $0 \leq i \leq 10$ contains the set of users that have already joined i sites. We perform the same classification for each set. Figure 3 shows the prediction results for different number of already joined sites. The figure also shows as a dashed line the majority class predictor for each set. We observe from this figure that the performance is the highest (97%) when users haven't joined any sites, and is decreased as users join more sites until 4 sites are joined. After which the performance starts to increase as more information about the user joining patterns becomes available to the algorithm.

Although the classifier performs the best when users haven't joined any sites, however, at this point the majority class prediction performs almost as well. The majority class in this case is the class of most popular sites. In other words, when users haven't joined any sites, they often just select the most popular sites; therefore, recommending these sites is most successful. We notice that as users join more sites, the effect of majority is reduced and when users have already joined 10 sites, the majority prediction is no different from random prediction ($25\% = \frac{1}{4}$). In other words, as users join more sites, peer pressure of joining popular sites is reduced and preference plays an important role. In this case, while the majority fails at predicting more than 30% correctly, our recommendation can perform as accurate as 60%.

Related Work and Conclusions

We have studied the user membership behavior across social media sites. We showed that user distribution across sites is a power-law distribution with an exponent of $\alpha = 2.34$. Using a weighted modularity measure, we computed the categories of sites that users join together. We show that users join some sites due to their popularity (YouTube, Twitter, etc.). There are also sites that users join all due to media (online radios/audio sharing/video sharing) and content (Social tagging/social bookmarking/social news) consumption purposes. The last category of sites that users join are new or relatively unknown sites. These are joined by early adopters who wish to explore and find new content. To evaluate these site selection patterns, we designed a site recommendation algorithm for users. We showed that while for users that are members of no site, recommending popular sites performs the best, users that have joined a few sites are more likely to select sites based on their preference.

Studying multiple networks has been the subject of a number of recent studies; see (Benevenuto et al. 2009; Magnani and Rossi 2011) for two such studies. The focus of these studies has been on how network dynamics and user behavior changes across networks irrespective of the users that these networks share or how behavior changes across networks after users join, irrespective of how these users select the sites in the first place. Our work is different from these studies as it analyzes individuals that are shared across networks, their distribution, and membership patterns.

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 perform better site recommendations. We consider this as a promising future direction for this work.

Acknowledgments

This work was supported, in part, by the Office of Naval Research grants: N000141110527 and N000141410095.

References

Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; and Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. In *SIGKDD*, 44–54. ACM.

BBC News. 2002. Turning into digital goldfish.

Benevenuto, F.; Rodrigues, T.; Cha, M.; and Almeida, V. 2009. Characterizing user behavior in online social networks. In *IMC*, 49–62. ACM.

Clauset, A.; Shalizi, C. R.; and Newman, M. E. 2009. Power-law distributions in empirical data. *SIAM review* 51(4):661–703.

Magnani, M., and Rossi, L. 2011. The ml-model for multi-layer social networks. In *ASONAM*, 5–12. IEEE.

Newman, M. E. 2006. Modularity and community structure in networks. *PNAS* 103(23):8577–8582.

Zafarani, R., and Liu, H. 2013. Connecting users across social media sites: A behavioral-modeling approach. In *SIGKDD*, 41–49.