Finding Friends on a New Site Using Minimum Information

Reza Zafarani and Huan Liu
Computer Science and Engineering
Arizona State University
Tempe, AZ 85281
{Reza, Huan.Liu}@asu.edu

Abstract

With the emergence of numerous social media sites, individuals, with their limited time, often face a dilemma of choosing a few sites over others. Users prefer more engaging sites, where they can find familiar faces such as friends, relatives, or colleagues. Link prediction methods help find friends using link or content information. Unfortunately, whenever users join any site, they have no friends or any content generated. In this case, sites have no chance other than recommending random influential users to individuals hoping that users by befriending them create sufficient information for link prediction techniques to recommend meaningful friends. In this study, by considering social forces that form friendships, namely, influence, homophily, and confounding, and by employing minimum information available for users, we demonstrate how one can significantly improve random predictions without link or content information. In addition, contrary to the common belief that similarity between individuals is the essence of forming friendships, we show that it is the similarity that one exhibits to the friends of another individual that plays a more decisive role in predicting their future friendship.

1 Introduction

With the rise of social media and the growth of modern technology, millions of sites are at our fingertips. With so many choices, our attention spans are decreasing rapidly. An average user spends less than a minute on an average site [1]. The problem becomes more challenging for commercial sites, especially for new sites that are desperately hoping to attract users and keeping them active. This lack of interest in users was clearly observed in the early years of social media sites such as Twitter or Facebook with around 60% of their users quitting within the first month [4].

As consumers of social media, we are constantly seeking "sticky" sites, that keep our attentions glued to the site by providing engaging material and more importantly, showing us a familiar face. The existence of familiar faces such as our friends, relatives, and our colleagues on one site, provides a sense of comfort, piques our interest on the site, and increases the likelihood of joining it [3]. By finding friends of individuals on social media sites, not only we increase users' engagement, but also improves user retention rates for sites, which could directly translate to more revenue for the social media site. So, *how can we find friends an of individuals*?

Finding or recommending friends is not a new problem [8]. It is a well-studied problem in social network analysis. Often, link or content information or a combination of both is used to predict and recommend friends to users.

When using link information, we use the current friends of an individual to recommend new friends. For instance, we find potential friends by finding individuals that are friend-of-a-friend. That is finding individuals that are 2 hops away in the friendship network. We can improve recommendations by recommending individuals that are more than two hops away in the friendship network. Unfortunately, recommending friends using link information fails when prior friends are unavailable. This can happen right after a user joins a new site, as a disconnected singleton in the friendship graph. Sites such as Twitter or LinkedIn, tackle this issue by asking users to provide access to their email contacts to help recommend friends. Aside from its security and privacy concerns, this clearly requires an extra effort from the user's side, and with the short attention span of a user, provides an opportunity for the user to abandon the social media site.

When using content information, friend recommendation techniques identify friends of an individual by identifying users that are highly similar to the individual in terms of the content that they generate. This content can be the profile information provided, the tweets, reviews, or blogs posted, or the products bought. However, right after a user joins a new site, he or she hasn't had the chance to complete their profile information or

exhibit any activity on the site.

In a sense, findings friends when no link or content information is available is a ubiquitous problem inherent to *all* social media sites and for each and every user, right after she joins the site. This problem is often referred to as the *cold start* problem.

The cold start problem has also been discussed in previous literature [10]; however, the approach to solving it often assumes that either link or content information is available. However, as we mentioned, when a user joins a new site, no link information (friends) or content information (bio, posts, etc.) is available; therefore, relying on either type of information may not be feasible. In practice, sites such as Twitter address this problem by recommending individuals that have many friends such as celebrities or political figures in the United States to newly-joined users. Some users may find these recommendation interesting, but it can be repelling to users that are from other countries or have limited knowledge of English. Ultimately, for a newly-joined user and without link or content information, finding friends in a site with one million members boils down to random recommendations of a few users from a *search space* of one million potential friends. Alas, recommending friends uniformly at random from this space is extremely unlikely to find any friends.

In this paper, we demonstrate a methodology to find friends on a new a social media site when link or content information is unavailable. Using social forces that form friendships, we demonstrate how one can employ minimum information from individuals to significantly reduce the set of potential friends in a social media site; hence, increasing the likelihood of finding friends. We discuss how this minimum information can increase friend finding performance sometimes by four orders of magnitude (Section 4). This can help sites introduce the very first few friends more accurately. Hence, increasing the chance for users to add friends, providing sufficient link information for more advanced link prediction techniques to recommend future friends.

Section 2 formally presents the problem of finding friends in social media sites with minimum information. Section 3 outlines how different social forces result in friendships and how one can utilize the outcome of these forces to tackle our problem. Section 4 outlines our experiments for finding friends. Section 5 reviews some related work. We conclude this work and provide future directions for research in Section 6.

2 Problem Statement

Consider a new site S with n users. When an individual joins S with no content or link information, the site has probability $p = 1/n$ to correctly recommend a single friend and a search space of n to search for that friend. If the user has no friends on S, no method is capable of finding any friends and all attempts to recommend potential friends from S fails. However, given the enormous size of current social media sites such as Twitter and Facebook, one can safely assume that the individual has some friends on the site.

Let set $U = \{u_1, u_2, \ldots, u_n\}$ represent the set of current users on site S and u_{n+1}, the newly-joined user. Consider a k-partitioning of current users $\pi(U)$,

$$\pi(U) = (X_1, X_2, \ldots, X_k), \tag{2.1}$$

$$\cup_{i=1}^{k} X_i = U, \tag{2.2}$$

$$X_i \cap X_j = \emptyset, \quad i \neq j. \tag{2.3}$$

To realistically model the problem in social media, and without loss of generality, we assume link information is available for current users $u_i \in U$, and unavailable for u_{n+1}. Assume link information is provided as an adjacency matrix $A \in \mathbb{R}^{n \times n}$, where $A_{i,j} = 1$ denotes that u_i is a friend of u_j; otherwise, $A_{i,j} = 0$. Consider *friendship matching function* $f : u_i \rightarrow X_j$ where $i \in [1, \infty)$ and $1 \leq j \leq k$ that matches user $u_i \in U \cup \{u_{n+1}\}$ (or later joining users $u_{n+2}, u_{n+3}, \ldots$) to a partition X_j, $1 \leq j \leq k$. We assume that the matched partition $f(u_i) = X_j$ is a partition in which it is highly likely to find friends for u_i. Thus, we denote partition X_j as the *friendship search space* for u_i. Let $M(X_j) = \{u_i | u_i \in U, f(u_i) = X_j\}$ denote the set of *matched users* from U to partition X_j. Since X_j is the set that most likely contains friends for all members of $M(X_j)$, we are implicitly assuming a level of similarity between members of $M(X_j)$. In our problem, we are seeking the friendship search space for u_{n+1}. For that, we need to determine $\pi(U)$ and function f. Even when both are known, assume X_j is the friendship search space for u_{n+1}. Since link information is unavailable for u_{n+1}, how can we verify if u_{n+1} has friends in X_j?

One approach is follow a training/testing framework in data mining and assume that the probability of u_{n+1} having friends in X_j can be approximated using other users matched to X_j: $M(X_j)$, for which we have link information. This probability ($P_f(X_j)$) approximates *link prediction accuracy* and is the fraction of matched users that have a friend in set X_j,

$$P_f(X_j) = \frac{|\{u_i | u_i \in M(X_j), \sum_{u_t \in X_j} A_{i,t} \geq 1\}|}{|M(X_j)|}. \tag{2.4}$$

For X_j, let $X_j^{Rand} \subseteq U$ denote a subset of equal size, $|X_j| = |X_j^{Rand}|$, where users in X_j^{Rand} are selected uniformly at random. Then, the probability that a user in $M(X_j)$ has a friend in X_j^{Rand} is

$$P_f(X_j^{Rand}) = |X_j^{Rand}|/|U|. \tag{2.5}$$

$P_f(X_j^{Rand})$ approximates random prediction accuracy for link prediction. Our goal in this study is to find friends by seeking partitions such as X_j, in which the probability of finding is much higher than random, i.e.,

$$\beta_{X_j} = \frac{P_f(X_j)}{P_f(X_j^{Rand})} > 1, \tag{2.6}$$

where β_{X_j} denotes the *significance ratio*[1], which quantifies the rate at which partition X_j increases the friend finding likelihood for members of $M(X_j)$. Note that the search space is reduced by $1/\beta_{X_j}$. Clearly, when no information is available, one cannot go beyond random: $\beta_{X_j} = 1$. The value for β_{X_j} is maximized when all users in $M(X_j)$ have at least one friend inside X_j. Thus, when sites such as Twitter recommend individuals with many friends (X_j ={celebrities or political figures}), they are providing a relaxed solutions to finding an optimal X_j.

The value of β_{X_j} can be deceiving, since for small values of $M(X_j)$, $P_f(X_j)$ can become large (see Eq.(2.4)); therefore, extremely larger than $P_f(X_j^{Rand})$. Furthermore, since u_{n+1} (and users joining later) can be matched to different partitions, one needs to compute the significance ratio for different partitions. Both issues can be addressed by computing the expected β for a partitioning[2] $\pi(U)$,

$$\mathbb{E}(\beta) = \sum_j \beta_{X_j} \frac{|M(X_j)|}{|U|}. \tag{2.7}$$

Thus, our goal is to find a partitioning of the users $\pi(U)$ and a friendship matching function f such that $\mathbb{E}(\beta) > 1$. To go beyond $\mathbb{E}(\beta) = 1$, we introduce the minimum information available on sites for users, excluding link and content information.

When users join new sites, the first step is to to create an account. Therefore, we consider the minimum amount of information available for an individual on a site to be her username. Usernames are alphanumeric strings or email addresses without which users are incapable of joining sites. To identify friends of an individual one can employ other information such as "first name+ last name", common friends, among other information. Unfortunately, there is no consistency in the availability of such information and we believe given those information, one should be able to use other methods and better find friends. These constraints directs us towards formulating our problem with usernames.

[1]Following the statistical convention of assuming $P_f(X_j^{Rand})$ as the null hypothesis, this ratio indicates how significant partition X_j is in predicting friends.

[2]In Section 3, we assume that $\cup_{i=1}^{k} M(X_i) = U$ and $M(X_i) \cap M(X_j) = \emptyset$; therefore, term $\sum_j |M(X_j)|$ is substituted with $|U|$.

Definition. ***Finding Friends with Minimum Information.*** *Given a set of usernames $U = \{u_1, u_2, \ldots, u_n\}$ and a friendship adjacency matrix $A \in \mathbb{R}^{n \times n}$, finding friends with minimum information can be achieved by finding a partitioning of U, $\pi(U) = (X_1, X_2, \ldots, X_k)$, and a friendship matching function f such that for $\pi(U)$, $\mathbb{E}(\beta) > 1$.*

So, the problem is reduced to finding a 1) partitioning of the usernames and a 2) matching of usernames to those partition such that $\mathbb{E}(\beta) > 1$. To tackle our problem, we analyze how friends are formed from a social science perspective.

3 Social Forces behind Friendships

In general, three major social forces result in friendships: 1) *homophily*; 2) *confounding*; and 3) *influence*. Homophily, best depicted in "birds of a feather flock together" is observed when *similar* individuals become friends. This similarity is often observed in terms of the interests of the individuals (e.g., their field of study) or their personal attributes that are unrelated to the environment they live in (e.g., gender). For instance, fans of the same movie director becoming friends is an example of friendships formed by homophily. Confounding is exhibited when friendships are formed due to the similarities in users formed by the environment they live in. Due to confounding friends are often in close proximity or speak the same language. Finally, Influence in friendships is observed when individuals form friendships due to an external factor such as the authority of an individual. For instance, befriending a public figure is due to influence.

Interestingly, signs of similarity among friends, also known as *assortative mixing*, are observed in friendships formed by all three social forces. In homophily, friends are similar in terms of non-environmental attributes such as their interests. In confounding, friends are similar in terms of their environmental attributes such as their mother tongue or location. In influence, after an individual befriends an influential, though the individual can be different from the influential in terms of the environmental or non-environmental attributes, but he or she often fits well within the **crowd** who has already befriended the influential. For instance, individuals who befriend a famous tennis player are often similar in terms of liking tennis. Thus, in influence, the user befriending the influential is *similar to the crowd that has befriended the influential.* Due to this assortative mixing, if a certain attribute of a user is known, say location, one can safely assume that 1) friends of the users are more likely to have the exact same attribute value (in case of homophily and

confounding) or 2) friends of the user are more likely to have friends that have the exact same attribute value (in case of influence). In either case, to find friends one should aim at predicting user attributes, and in our situation, from usernames.

We believe that due to unique personal attributes, individuals exhibit certain behaviors. These behaviors are non-random and therefore, create information redundancies [13]. These information redundancies can be captured in terms of data features in their usernames. For instance, we expect individuals who speak the same language to have statistical language patterns observable in their usernames. Following the tradition in machine learning and data mining research, we employ supervised learning techniques to predict personal attributes of users solely from their usernames. For each social force, we select a corresponding user attribute for prediction that can best demonstrate the effect of friendships formed by that force. Next, we elaborate how specific attributes of users are selected for each social force to be predicted from their usernames.

3.1 Predicting Individual Attributes As discussed friendships are formed by three general social forces: homophily, confounding, and influence. Our goal is to predict user attributes that are observable in usernames and represent each social force. Note that one can try to predict many attributes from usernames for different forces and this should help better understand each social force and find more friends. We leave this as part of our future work. Our goal here is to demonstrate how simple user attributes that represent each social force can be predicted using only usernames. Later on in Section 4, we measure how these attributes help better find friends and we measure the effect of each social force on predicting friendships.

3.1.1 Homophily-based Friendships Homophily is observed when individuals befriends each other due to their similarities in non-environmental user attributes. A major non-environmental user attribute that is shown to introduce friendships is the age of the individual. One often observes that individuals in the same age range more frequently befriend each other. This has been observed in numerous recent studies [11, 9]. For instance, Ugander et al. [11] noticed that younger individuals have less diversity in the age of their friends while older ones have a much wider range. Among the many attribute that result in homophily-based friendships, we select age due to to its strong affect on friendships. If the ages of individuals, represented as usernames are predicted, one expects usernames in the same age range to have higher friendship likelihoods.

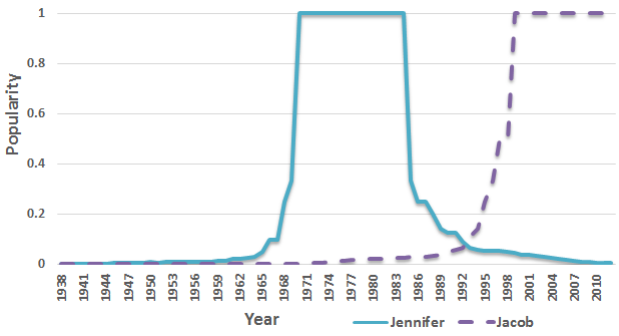


Figure 1: Popularity of first names: *Jennifer* and *Jacob* over time. Higher values depict more popularity.

By predicting the ages of usernames in the network, we are partitioning the network into different sets, each set representing an age range. Now, for a new username, we can predict the age range of it, and we expect the username to have a higher likelihood to be connected to the users in the partition with the same age range. In other words, $\pi(U)$ becomes the partitions of different age ranges, $f(u_i) = X_j$, where the predicted age range for u_i is the same as all members of X_j, and $M(X_i) = X_i$ meaning that matched users are within the partition itself. The question is, how do we predict the age of an individual that owns a username?

An analysis of US social security records[3] for birth names since 1879 shows us that the frequency of different names change over time. For instance, in Figure 1, we depict the popularity of first names: *Jennifer* and *Jacob* over time. For each year, the popularity of the first name is shown on a scale of [0,1]. To compute popularity, the frequency of the names is measured in a year, and then, the rank of the name is computed among the frequency of all names born in that year. The inverse of the rank is considered as the popularity. Jennifer was the most popular female name between [1970-1984] whereas Jacob has been the most popular male name for the last 13 years [1991-2012][4]. Similar patterns can be observed for different English and non-English names given the diversity of the US population. This leads us to believe that given a name, one can provide an estimation of a likely age.

Personal attributes such as names have been shown recently by Zafarani and Liu [13, 12] to influence the usernames, measured by alphabet distributions that change over time. For instance, in our example, the probability of observing double n's in *Jen**n**ifer* decreased over time and the probability of observing c and

[3] http://www.ssa.gov/oact/babynames/
[4] Data for 2013 is still not available.

b increased over time due to the popularity of ***Jacob***. Similarly, we observe that the n-grams in names change over time. Furthermore, individuals of different age ranges have different vocabularies demonstrated in their usernames [13]. Thus, one can employ statistical language patterns to estimate age of the individuals who own usernames.

3.1.2 Confounding-based Friendships Among the many attributes that describe the environment that the users are living, we select two of the most prominent attributes: their language and location. Similar to the age of individuals, we expect users living in close proximity or sharing the same language to have a higher chance of becoming friends. Similar to the age attribute, $\pi(U)$ becomes the partitions of different locations (or languages), $f(u_i)$ matches u_i to the partition X_j where are members of X_j are in the same location as u_i, and $M(X_i) = X_i$, meaning that matched users are within the partition itself.

The language of individuals can significantly impact their chosen usernames. The language patterns can be easily observed both in the alphabet distribution as well as the n-grams of the username. For instance, while letter x is common when a Chinese speaker selects a username, it is rarely used by an Arabic speaker, since no Arabic word transcribed in English contains letter x. Similarly, excessive use of 'i' in languages such as Persian or Tajik [5, 7], can be easily detected in usernames.

Similarly, individuals from specific location often have tendencies to utilize words or statistical patterns that are only observable in those regions. While a native of Zambia, may use word `Kalambo`, referring to a waterfall in Zambia, it is highly unlikely for users from elsewhere to include this word.

Thus, to predict the location and language of the individuals one can identifying statistically meaningful alphabetical patterns in their usernames.

3.1.3 Influence-based Friendships When friendships are formed due to influence, we are assuming influential users are attracting friends. In this scenario, we can partition the users attracting other in terms of the types of friends they are attracting and compare each partition with the user for which we are searching for friends. In general, we believe the factor that is deciding in becoming a member of the crowd that has befriended an influential is how the user fits in that crowd. We assume that a user fits in a crowd when at least one member of the crowd is similar to the user in terms of some attribute (environmental/non-environmental). We use all three attributes predicted so far: age, location, and language to predict this similarity. Here, $f(u_i)$ matches u_i to a partition X_j where each member of X_j has a friend with the same language, location, or age as u_i.

4 Experiments

The friendship space reduction is systematically evaluated in this section. We would like to verify if assortative mixing is successful in finding friends for each one of the social forces represented by their predicted attributes. Before we present our experiments, we detail how experimental data is collected.

4.1 Data Preparation To analyze the friendships, we collected a friendship graph of 135 million friendships from social media site Reddit. These friendships are among 1.6 million users. For each friendship in this graph, we have the two usernames that are connected. We also collected separate datasets for predicting age and location of usernames. All datasets employed in this study are shareable for research purposes.

4.1.1 Age Dataset To predict age and to remove any bias associated with the usernames in Reddit, we collected a set 226,588 usernames from LiveJournal. In LiveJournal, users can list their age. Among these users, 82,011 users have listed their age. This formed our training dataset for age prediction. The usernames in this dataset were vectorized using their alphabet distribution and frequent letter bigrams and their weights were normalized using TF-IDF. The ages were also divided into ten categories using an equal frequency binning and used as labels of this dataset. The age ranges in years are: $[0, 21.9]$, $[22, 23)$, $[23, 25)$, $[25, 26.5)$, $[26.5, 28)$, $[28, 30)$, $[30, 33)$, $[33, 36)$, $[36, 42)$, $[42, \infty)$.

4.1.2 Location Dataset Similar to the age dataset, to remove bias for location prediction, we collected a dataset from Twitter. On Twitter, individual tweets can be geo-located; that is, users carrying gps-enabled devices can report their location with their tweets, which includes their usernames. The location is reported in (latitude,longitude) format. From Twitter, we collected a set of 36 million geo-located usernames with their latitudes and longitudes. Using a shapefile of all country borders and reverse geocoding, we determined the country for each username. Clearly, some countries have more geo-located tweets than others. To account for this imbalance, we clustered our dataset of latitudes and longitudes with k-means clustering.

For countries with less than 1,000 usernames we considered the whole country as one cluster. For all others, we clustered the geographical coordinates within the country using k-means with different k values until

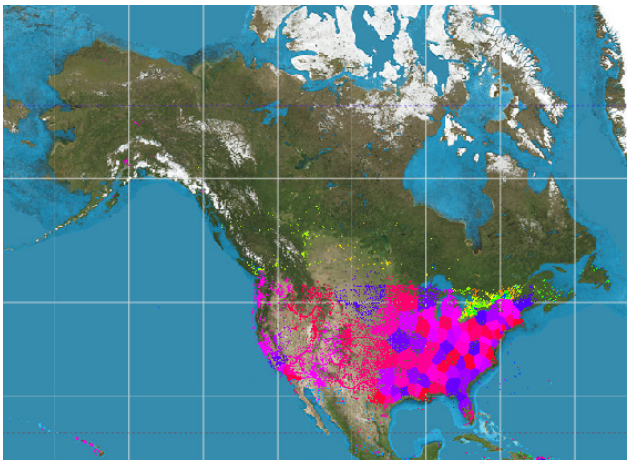


Figure 2: Usernames Clustered based on Location for the United States. Colors Represent Cluster Labels.

the obtained clusters had small enough radius. A recent study on Facebook [11] shows that users are more likely to befriend individuals that are within their 50 miles distance; thus, we ensured that the distance between any two members of the same cluster is close this value. In our dataset, we found that by finding around 395 clusters, the clusters become well-balanced in size and small in radius across countries. The clustering of the usernames from the United States, including Alaska and Hawaii, is shown in Figure 2.

Although some clusters were still smaller than others, for most clusters, the difference is negligible, with the average datapoint distance to the cluster centroid being ≈ 36 miles. Since users in the same cluster are geographically close, we expect these users to have higher friendship likelihood. In this dataset, we use the cluster label as the class label for our training. Similar to our age dataset, the usernames are vectorized using their alphabet distribution and frequent letter bigrams and their weights are normalized using TF-IDF.

4.1.3 Preparing Age, Location, and Language Predictors

We discuss how we train different classifiers to predict age, location, and language. Note that we are agnostic to the performance of these classifiers and our goal is to train a classifier that can reasonably predict each attribute. This is due to our goal to demonstrate the feasibility of finding friends by training such classifiers. Clearly, if our classifiers are capable of helping find friends, further improvements can improve the performance even more. We leave improving classifiers as future line of research.

Predicting Language of Usernames. Since usernames are often transliterated in Latin alphabet, one can only predict the language of usernames for languages that employ Latin alphabets. We train an n-gram statistical language detector [6] over the European Parliament Proceedings Parallel Corpus[5], which consists of text in 21 European languages (*Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene,* and *Swedish*) from 1996-2006 with more than 40 million words per language. The trained model can detects a username's language by decomposing it into different n-grams.

Predicting the Age for Usernames. Given our prepared dataset for the age. We trained a regularized logistic regression model that is able to predict the age of a username by decomposing it into n-grams. The model is used later to predict ages for other usernames.

Predicting Location of Usernames. As discussed, the location dataset was clustered based on latitude-longitude values and then cluster labels were used as class labels. We trained a regularized logistic regression model for this dataset where the model is capable of detecting the location of the username as one of the 395 classes that represent different locations.

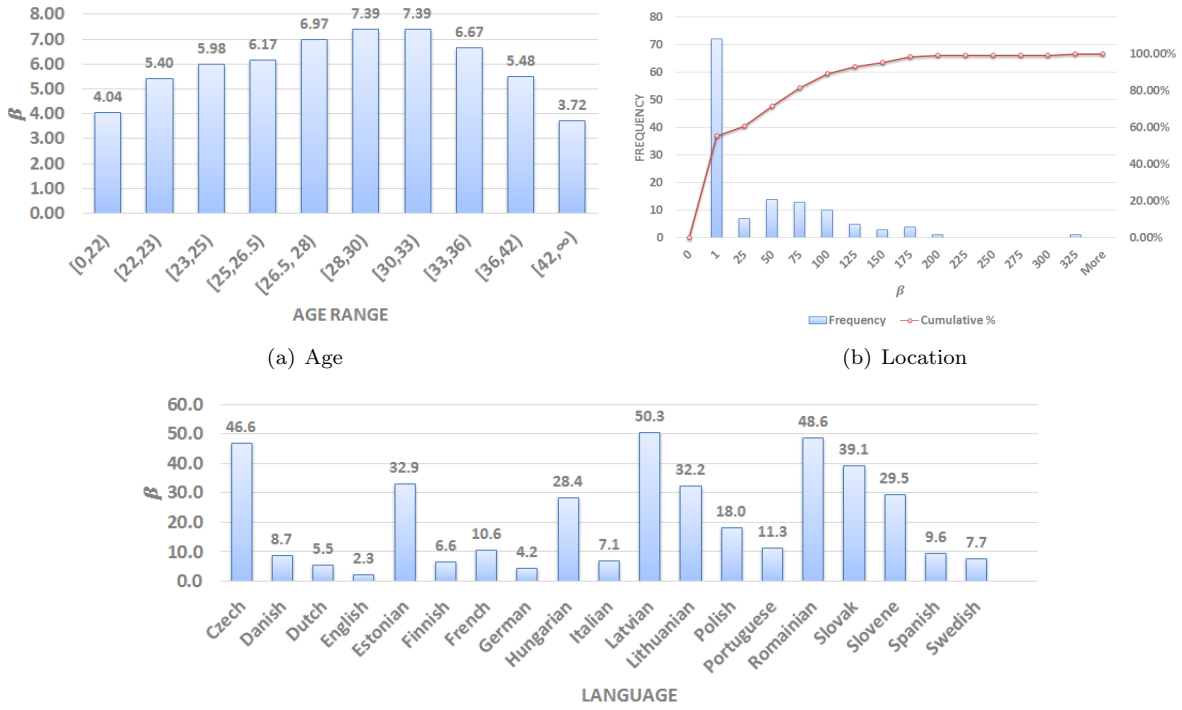
4.2 Measuring Significance Ratios

Given our trained classifiers, we perform age, location, and language prediction for all 1.6 millions users. Then, for attributes representing each social social force, we measure the significance ratio. For homophily and confounding, we measure significance ratios by measuring how many friends are of the same age, location, or language. For influence, for user u_i and user u_j (represented using usernames), we measure how username u_i fits among the friends of u_j. We perform this separately for each of the three predicted attributes. In our experiments, we assume user u_i fits in friends of u_j, if at least one individual among friends of u_j has the same attribute value (age, location, or language) as u_i.

4.2.1 Homophily Significance

Among the set of 135 million friendship users we measure significance ratios for all age categories in our dataset: $[0, 21.9]$, $[22, 23)$, $[23, 25)$, $[25, 26.5)$, $[26.5, 28)$, $[28, 30)$, $[30, 33)$, $[33, 36)$, $[36, 42)$, $[42, \infty)$. The significance ratios are plotted in Figure 3(a). As shown in the figure, for all categories $\beta > 1$. This means that for example, when the predicted age of a username is between $[28 - 33]$, by recommending only other usernames where their ages are predicted to be in $[28 - 33]$, we are 7 times more

[5] http://www.statmt.org/europarl/



(a) Age

(b) Location

(c) Language

Figure 3: Significance Ratios (β) for Different Attributes

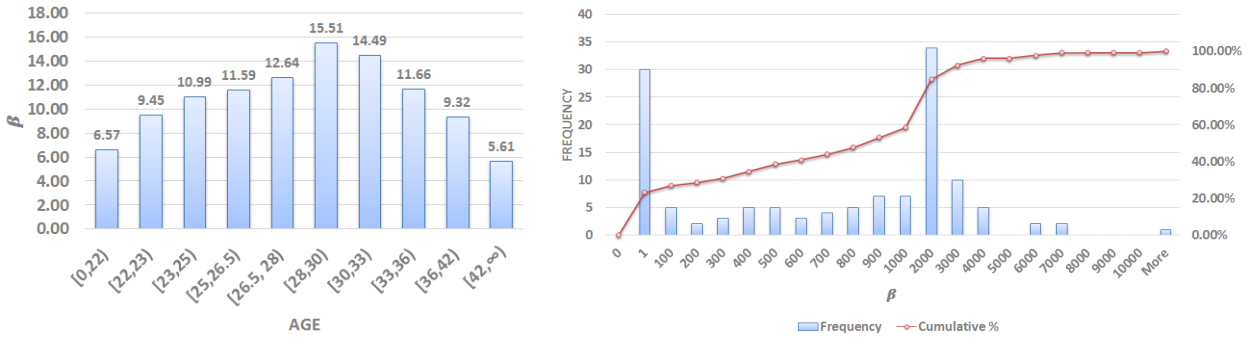
accurate than randomly finding a friend. Note the significance of this result, compared to state of the art link prediction techniques that perform on average 2.4-54.4 times better than random prediction [8]; however, with access to *link information*. Our technique has no access to link information for the individual for which we are finding friends.

4.2.2 Confounding Significance Similarly, we measure the significance ratios for different languages. We observe that for all languages $\beta > 1$. More importantly, we observe that when the language is detected as English, then β is minimum among all languages. This has two reasons. First, the majority of usernames are in English; therefore, conveying less information about friends. Secondly, lower assortative mixing is observed among English users, as English is widely spoken across the globe and there is less likelihood for these speakers to befriend each other. In direct contrast are eastern European languages such as Romanian ($\beta = 48.6$) or more commonly spoken languages such as French ($\beta = 10.6$) that significantly improve friend finding performance.

We also measure the significance ratios for the location of usernames. Due to the large number of locations, we plot the histogram and the cumulative distribution (red line) of β values in Figure 3(b).

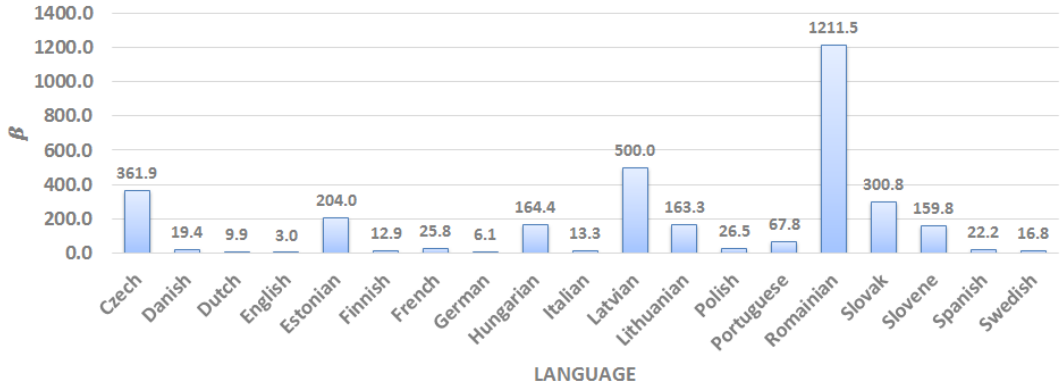
As shown in the figure, for more than 55% of locations we cannot predict any better than random. At the same time, for some predicted locations one can achieve as much as $\beta \approx 325$. After further investigation, we found that for the locations where $\beta = 1$, either the radius of the location cluster was larger than 50 miles or the size of the username cluster was small (few training instances). This in particular happens for countries where not many usernames are in our dataset. Thus, to better understand if there is any significance with respect to location, as well as other attributes, one needs to compute the expected value $\mathbb{E}(\beta)$. We will measure the expected values later in section 4.3 where we compare different social forces in terms of friend finding performance.

4.2.3 Influence Significance We measure significance ratio for influence using age, location, and language. These ratios are demonstrated in Figures 4. Comparing Figure 4 to Figure 3, we observe that in general finding friends based on influence (similarity to the friends of an individual) is much easier compared to homophily or confounding. On average, when using influence, and using attribute age, the friend finding performance is improved by a factor of 1.79. Similarly,



(a) Age

(b) Location



(c) Language

Figure 4: Influence Significance Ratios (β) for Different Parameters

Table 1: Expected Improvement in Finding Friends over Random Predictions ($\mathbb{E}(\beta)$) for Different Social Forces.

Friend Finding Technique	$\mathbb{E}(\beta)$
Homophily - Age	5.49
Confounding - Location	6.19
Confounding - Language	5.19
Influence - Age	9.79
Influence - Language	16.29
Influence - Location	**31.04**

it is improved by a factor 5.14 when using the language attribute and a factor of 11.72 times when considering locations. Hence, it seems that users prefer befriending individuals that have friends in their region over individuals who have friends talking their language or are of the same age. To further analyze each social force we measure the expected improvements in finding friends for each social force next.

4.3 Comparison between Social Forces As discussed in Section 2, the significance ratio at times can become deceiving. To mitigate this issue, we compute the expected β for homophily (age attribute), confounding (language or location attribute), and influence (for age, location, and language). The results are available in Table 1, showing an expected improvement factor between [5.49-31.04]. We observe in the figure that though all forces can help find friends, influence-based friendships, by at most a factor of 6, best find friends compared to the other social forces. Contrary to the common belief that similarity between users is the gist of forming friendships, this suggests that individual have far more tendencies to befriend a potential user when they feel welcomed in the crowd of friends of the potential user. We observe no significant difference between homophily and confounding in finding friends.

5 Related Work

To the best of our knowledge, our study is the first to help find friends when no link or content information is available. However, one can find similar unsupervised link prediction studies in the existence of link or content information that are applicable in our case.

Assuming usernames are content generated by users, one can compute the similarity between individ-

uals and the similarity between their friends. In this case, well-established link prediction methods that use node similarity or neighborhood similarity such as the common neighbors [8], Adamic-Adar [2], Jaccard's Coefficient [8], or preferential attachment[8] are applicable. Note that when using contents generated by users, it is common to assume large collections of documents, with thousands of words, available for each user, whereas for usernames, the information available is limited to one word. Our technique, employs the knowledge of how social forces influence friendships and additional information such as age, language, and location that represent these social forces to reduce friendship search space, helping better predict future friends.

6 Conclusions and Future Work

In this paper, we propose an approach for finding friends when link or content information is unavailable. This problem is ubiquitous to all social media sites since when a user joins a new site, he or she has no friends or has not generated any content. Under these constraints, sites are often forced to recommend randomly chosen influential friends, hoping that users by adding these friends, create sufficient information for link prediction techniques for further recommendations.

Friendships in social media are often formed due to three social forces: homophily, confounding, and influence. We show how minimal information available on all social media sites (usernames) can be employed to determine friendships due to these forces. In particular, we employed usernames to predict personal attributes such as age, location, and language that in turn can be used to find friends and measure the effect of each social force. Our empirical results show the advantages of this principled approach by improving friend finding performance by an expected factor of 5.49-31.04 over random prediction. This is comparable to the state of the art link-prediction techniques that perform 2.4-54.4 times better than random prediction [8]. Our results also show that while by employing each social force, one can improve friend finding performance at least by a factor of 5.49, influence social force can help best find friends. This suggests that individuals have more tendency to befriends others with similar friends (influence), than those who are more similar to them (homophily) or live in a common environment (confounding). Our results show an improvement of at least a factor of 6 over random predictions when link or content information is unavailable; hence, increasing the likelihood of retaining users and keeping them engaged. Note that using our method personalized recommendations are performed since for example, users identified as French are more likely to be recommended French users.

Our work opens the door to many interesting applications. Studying addition of other information or analyzing how combining this approach with traditional link prediction can further improve the performance of link prediction are examples of the many areas that can benefit from the results of this study. Future work also includes analyzing these possibilities and discovering how these social forces can be combined to further improve friend finding performance. While we demonstrated that all social forces are helpful in finding friends, the comparison of forces can be influenced by the performance of the classifiers. We leave verifying our findings with labeled data in which age, location, or language is known as another part of our future work.

ACKNOWLEDGMENTS

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

References

[1] Turning into digital goldfish. *BBC News*, 2002.

[2] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[3] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD*, pages 44–54. ACM, 2006.

[4] P. Cashmore. 60% of Twitter Users Quit Within the First Month. `http://on.mash.to/zVwKb`, 2009.

[5] D. Cowan. *An Introduction to Modern Literary Arabic*, volume 240. Cambridge University Press, 1958.

[6] T. Dunning. *Statistical Identification of Language*. CR Lab, New Mexico State University, 1994.

[7] C.A. Ferguson. Word Stress in Persian. *Language*, 33(2):123–135, 1957.

[8] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[9] Mark Newman. *Networks: an introduction*. Oxford University Press, 2009.

[10] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. Methods and metrics for cold-start recommendations. In *Proceeding of SIGIR*, pages 253–260. ACM, 2002.

[11] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[12] Reza Zafarani and Huan Liu. Connecting corresponding identities across communities. In *ICWSM*, 2009.

[13] Reza Zafarani and Huan Liu. Connecting users across social media sites: A behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD*, pages 41–49. ACM, 2013.