I Seek You: Searching and Matching Individuals In Social Networks

Marti Motoyama and George Varghese
University of California, San Diego
9500 Gilman Drive
La Jolla, California 92093
{mmotoyam,varghese}@cs.ucsd.edu

ABSTRACT

The first task any individual faces after joining an online social network (OSN) is locating friends that are present on that particular site. Most OSNs offer some variation of a tool that imports email contact lists to facilitate the task of finding one's friends. However, given that OSNs attempt to reconnect individuals with past acquaintances, one might not have access to the email address for a long lost friend. Furthermore, people tend to utilize a number of aliases online, meaning that an email address cannot always be used to reliably find a friend. Thus, new members must still manually search for friends based on a number of biographical attributes, such as gender, age, hometown, etc. It is not clear, however, what attributes are useful for conducting the search. Even after the search has been performed, the person performing the search might be left with a number of candidate profiles. In this paper, we develop a system for searching and matching individuals in OSNs. We evaluate the efficacy of our person matching techniques by measuring the overlap between social networks, and comparing our results to those published by compete.com. We then look at several interesting properties of overlapping profiles in both networks.

Categories and Subject Descriptors

C.2.4 [**Distributed Systems**]: Distributed applications; H.3.3 [**Information Search and Retrieval**]: Search process; H.3.5 [**Online Information Services**]: Web-based services

General Terms

Design, Measurement

Keywords

Social Networks, Crawling, Entity Resolution

WIDM'09, November 2, 2009, Hong Kong, China.

1. INTRODUCTION

Online Social Networks (OSN) continue to grow at phenomenal rates. Facebook (FB), for example, has increased to over 200 million active users in recent months [13]. As more and more individuals join OSNs, the task of locating former colleagues and peers becomes increasingly difficult. Name collisions, unlisted personal details, and variations in biographical attributes all contribute to the difficulty of discovering friends. OSNs attempt to alleviate the problem by offering a tool that imports email addresses. Since OSNs require that all members specify an email address during the registration process, the tool presumably searches for individuals using the email addresses present on the contact list. However, prior studies have shown that individuals tend to have multiple aliases [6]. As such, the address that a person uses to exchange emails may not be the same as the address used to register with the OSN. Furthermore, a person may not know the email address for the individual he or she is seeking, since one purpose of an OSN is to reconnect members with out-of-touch acquaintances. Thus, people need to rely on other biographical features to locate their friends, resulting in a search process that is much less precise.

The goal of this research is to provide insight into what search "keys" (e.g., email address, name, age, gender) are useful for locating individuals on social networks. We investigate which criteria are more effective at locating people, in the face of search attributes that are potentially widely shared. For example, there are over 144,000 individuals with the name "John Smith" on Facebook. We present a system that searches for people in OSNs, issuing queries based on a number of profile attributes (including name, age, geographical location, etc.). The searches we conduct generally yield sets of candidate profiles. In order to determine whether the search succeeded, we trained a classifier using boosting to identify whether a match exists in the set.

To evaluate the effectiveness of our searching and matching techniques, we randomly select profiles from FB and MySpace(MS). If the profile originated from FB, we search for the person on MS, and vice versa. This seeding and searching performed on different social networks roughly models the uncertainties that people face while conducting searches. We expect that the number of individuals we locate should approximately match the overlap numbers reported in the 2007 study conducted by compete.com [3], adjusted for the growth that both OSNs have experienced since 2007. The results of our evaluation allow us to observe several interesting facets of overlapping profiles.

The main contributions of this paper are:

- **Methodology:** We describe our methodology for searching and matching individuals across OSNs.
- **Results:** We present an analysis of the overlapping profiles. We look at several properties that the profiles exhibit, including conflicting security policies, active overlap, etc.

This paper is organized as follows: Section 2 provides a brief introduction to the unique features of the two major OSNs we targeted for our study. We then discuss motivations for our study in Section 3, and describe our methodology in both Section 4 and Section 5. Section 6 presents our results, Section 7 discusses related work, and Section 8 concludes.

2. BACKGROUND

Participants in the Facebook and MySpace communities create profiles where they specify personal details such as names, birth dates, locations, interests, colleges attended, etc. on their profile pages. The purpose of an OSN is to facilitate the formation of "friend" relationships within its user population. Subsequently, members who are friends on an OSN can share information, exchange multimedia, and communicate asynchronously. To locate acquaintances, each OSN provides a means of searching for individuals based on personal attributes such as email addresses, names, ages, etc.

In order to evaluate our system, we randomly select profiles from one OSN, and search for the individuals on a separate OSN using the attributes we extract from their profiles. We now describe the privacy and querying models for each of the aforementioned OSNs, which we must consider when formulating a measurement architecture.

Facebook

Facebook is currently one of the largest and most popular social networks in the world. FB has a fairly stringent security model, as users are only able to view profiles for individuals in their own "networks." Networks are centered around geographic regions, companies, and educational institutions. A member of FB can join only *one* geographic network, but can participate in any number of education/employment networks, as long as they have valid email accounts associated with the institution. For example, if a user has access to an email account for "john@foobar.edu", he can join the "foobar" university network.

FB allows users to customize their security settings. The default security configuration allows one to view a person's profile only if the two individuals (the viewer and viewee) share a common network. Email addresses on Facebook are encoded as PNG images, and Facebook has some means of enforcing that individuals use their real names during the account creation process. Lastly, FB provides a robust search engine, allowing one to issue a query for an individual based on a variety of profile attributes. In particular, FB allows searching by one's company or educational institution, which are particularly discriminating features when searching.

MySpace

MySpace is another popular social network, but lacks the notion of a "closed network" as defined by FB. MySpace allows individuals to specify whether their profile is accessible to only their friends or to arbitrary users. Despite this ability, MySpace community members tend to have public profiles. A fairly significant impediment to our study when dealing with MySpace profiles is that individuals typically *do not* list their real names, although in recent months, the site has been pushing users to reveal this information.

MySpace allows for searching based primarily on an individual's geographic location. MySpace permits searching by age, which proves valuable since Facebook users generally specify their birth dates.

3. MOTIVATION

We motivate our study by first investigating the effectiveness of *only* utilizing email addresses to locate friends. Next, we present several applications of our system and the data we can collect.

3.1 Email Search Effectiveness

We collected 953 random Facebook profiles on May 12th, 2009 from the 34 most popular geographic networks on FB. Each of the 953 profiles contained at least one email address image, which we ran through an OCR engine. We searched for these email addresses on MySpace and discovered 173 overlapping profiles. Thus, our preliminary analysis leads us to conclude that 18% of FB overlaps with MS.

However, compete.com's [3] study in November of 2007 revealed that 64% of Facebook overlaps with MySpace. We must adjust this percentage according to recent size estimates of Facebook and MySpace. A recent study [15] shows that Facebook grew from 92,754,000 unique visitors in Nov. 2007 to 200,189,000 in Nov. 2008. During the same time period, MySpace grew from 104,473,000 to 120,691,000.

If 64% of Facebook users in 2007 were also present on MySpace, the absolute number of MySpace users in Facebook was $0.64 * 92,754,000 \approx 60$ million in Nov. 2007. Assume that *none* of the 107,435,000 new users who joined Facebook in 2008 also joined MySpace. The percentage overlap in November 2008 would be approximately 60m/200m, which is 30%. Now, assume that *all* ≈ 16 million new MySpace users also joined Facebook, then the percentage overlap in Nov. 2008 would be at most 76m/200m (38%).

In either case, the 18% overlap we observed during our preliminary analysis is lower than we expect. Hence, we conclude that using only email addresses is insufficient for finding individuals in OSNs. Thus, we investigate other methods for locating people, in addition to using their email addresses.

3.2 Online Identity Maintenance

As a consequence of building our searching and matching infrastructure, we can obtain data that provides us with insight about how individuals maintain their online identities. If we can identify users across social network boundaries, then we can observe how individuals have reacted to the proliferation of the OSN market (Wikipedia lists over 50 OSNs). For example, do users join one OSN to ease the process of maintaining their online identities, or do they create accounts in many OSNs to achieve greater accessibility? A user that is part of multiple OSNs must synchronize (or maintain) his online persona on each OSN he is a part of.

These questions become especially pertinent with the emergence of online services to ease the process of synchronizing online personas [9]. Clearly, there seems to be interest in

	Search Capabilities									
Social Network	Name	Email	Education	Employment	Gender	Age	Country	City	Hometown	Zip
MySpace	Y	Y	N	N	Y	Y	Y	Y	N	Y
Facebook	Y	Y	Y	Y	Y	N	Y	Y	Y	Y

Table 1: Search capabilities for Facebook and MySpace

addressing the tradeoff between maintainability and accessibility. With the data we obtain, we can provide *quantitative* insight into the extent of this need. Do users of OSNs generally have multiple accounts? If they do, is it because they have disjoint sets of friends in each network? Also, do users successfully maintain their profiles in all OSNs they participate in?

While quantitative answers are of interest to social scientists and to social network aggregators, they are also of interest to OSN providers. In the future, two or more OSNs may merge in some fashion analogous to the way ISPs peer. Intuitively, an OSN should peer with another OSN that has as *disjoint* a set of users as possible. Once again, this requires quantitative information about the degree of overlap between disparate OSNs.

4. SEARCHING METHODOLOGY

When we encounter a profile on Facebook and MySpace, we download several HTML pages linked to the individual. The pages we extract contain useful information for identifying a particular person. The list of pages includes: profile, comments, photo albums, email images, applications, blogs, and friends.

We issue search queries using wrappers we created for the people finding search engines offered on FB and MS. We obtain data for our search queries by extracting personal details from the aforementioned profile pages. Of course, the information we pull out from the profiles is driven by the search capabilities of the social networking sites. Table 1 lists the attributes we search for on MySpace and Facebook. Both sites permit searching by email and by a number of biographical attributes (name, age, gender, etc.).

When we search based on personal information, we do not search all the possible permutations of the attributes. Instead, we order the personal details according to increasing restrictiveness (we call these orderings "search chains"). We specify more and more details to reduce the number of matching candidates. For example, suppose we know that person p is from the United States, and p's age is 29. First, we search only for p's name. If the number of search results is over threshold T (T is arbitrarily chosen as 10 during our experiments), we search for p again, specifying p's name and "United States" during the subsequent query. Lastly, we include p's age (in addition to the aforementioned attributes) during the final search attempt. We do not consider the ranking of the T possible profiles returned from the search engine during our classification, since we are unsure how FB or MS rank the search results.

Facebook offers a larger toolset for conducting person searches compared to MySpace (see Table 1). In particular, searches may be conducted based on a person's educational and employment history. In order to conduct these searches, we must query Facebook's AJAX servers to find the identification number and standardized name that corresponds to the company or educational institution. For example, suppose we are attempting to search for person p, and we know that p attends "UCSD." Facebook resolves the name "UCSD" to an internal identification number, which we must specify on the search form.

At the conclusion of the search, we are generally left with multiple candidate profiles that may match the source entity.To deal with this issue, we use a classifier to check whether a match exists among the candidate set, a process that we discuss in 5.2. However, since we require training data to build the classifier, we first elaborate on the architecture we employed for searching and collecting profiles.

5. DATA COLLECTION

The data collection system we used to perform the seeding and searching is comprised of three components: *supervisors, vaults*, and *workers*.

Workers: We instantiate a set of workers for each network N involved in our evaluation. Workers are synonymous with crawlers and represent our interface to the OSNs. The workers are responsible for both collecting *seeds* and processing any search queries within their transaction queues. A seed is simply a representative user profile from network N, who we later search for on a separate social network. The workers choose seed profiles that contain real names, since attempting to search FB without a name is nearly impossible. For FB, this is not an issue, since FB enforces the policy that all participants use real names. MS recently began listing real names for users, so the workers select users who reveal their real names. The workers upload the content of the profile pages to a database. Step 1 in Figure 1 depicts how we ascertain the seeds S.

Supervisors: We implemented supervisors that mine the content of the profiles in the seed set S, extracting the searchable attributes discussed earlier (names, emails, age, gender, education, etc.). Recall that these attributes are useful to resolve ambiguities during the search process, especially when several users exist with the same name (e.g., "John Doe"). This is shown as Step 2a in Figure 1 where the supervisor augments the "John Doe" profile with the country Australia by consulting Yahoo Maps.

Once the supervisor has extracted the relevant information from a profile, the supervisor selects the worker holding the credentials (see Section 5.1 for account details) associated with the geographic location linked to the profile. Recall that FB enforces the security policy that a person viewing a *random* user's profile must "share" a network (in our case, geographic) with that user. Furthermore, recall that we extract seeds from specific geographic locations, so we can always link the extracted profile to a particular country/city, even if the person does not specify a location. This is shown as Step 2b in Figure 1, where a search request for MySpace user John Doe is serviced by a Facebook worker attached to the Australian geographic network. The search process will return a set of candidate matches $Candidates(S)$ for each seed in S.

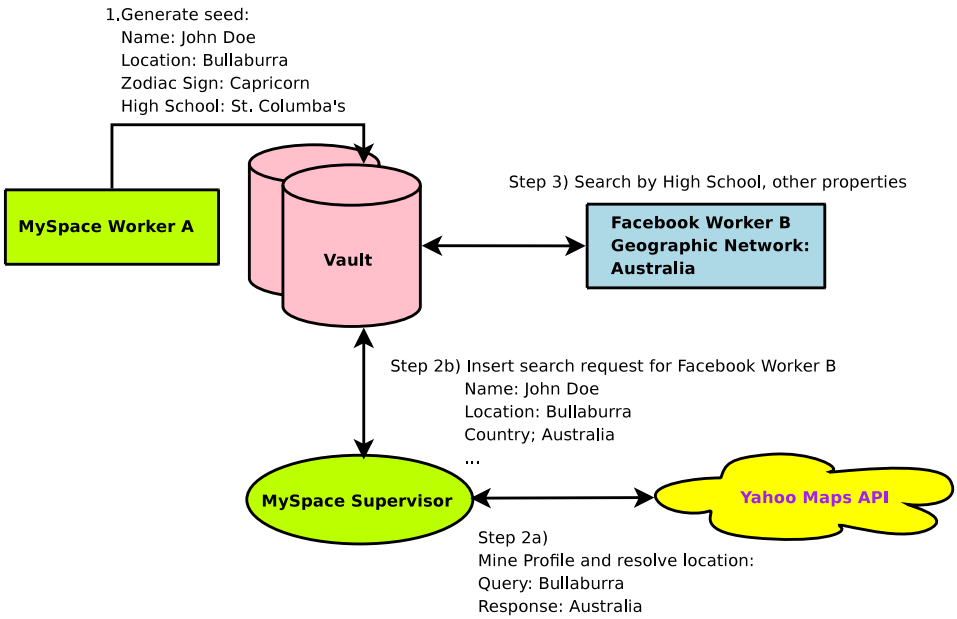

Figure 1: Our crawling infrastructure consists of Vaults, Supervisors, and Workers. Vaults are responsible for storing information and inputting/retrieving work requests. Supervisors mine profiles and generate search stubs. Workers generate seed profiles and process search requests.

Vaults: All profiles are stored in a database. Vaults arbitrate access to the database and wrap SQL operations. Our database maintains not only the data from each profile, but is responsible for housing the search requests. Using a database for storing transactions is key, as we can track requests across system restarts.

In a second analysis phase (not shown in Figure 1, which only represents the data collection phase), we run our classifier on the set of candidate matching profiles $Candidates(S)$ to determine whether profile S indeed matches with any candidate. We now elaborate on the challenges we faced when constructing this data collection system.

5.1 Implementation Challenges and Details

We faced the following challenges in collecting our data:

- **Coordinating Distributed Workers:** First, to avoid crawling restrictions imposed by OSNs and to achieve scalability, we designed our system in a distributed manner. Previous versions of OSN crawlers appear to scale by using *independent* crawlers — each crawler acts on a portion of the social graph without synchronizing with other crawlers. By contrast, our workers need to synchronize and share information to conduct the profile searches described above. As discussed earlier, FB only allows a worker to view a profile if the worker shares a network with the person of interest. We need to direct a search request to the appropriate machine hosting the worker for a particular location. We deployed the workers, supervisors and vaults on local campus virtual machines (VMs) using Usher [10]. The VMs hosting the workers were instantiated with publicly addressable IP addresses. We used the Ruby Mechanize/Hpricot packages to code all of our modules and classes.

- **Extracting Attributes from Profiles:** Recall that to find candidate matches, we need to extract attributes from a profile. Names by themselves are insufficient, because many individuals share the same name, thus producing a number of name collisions during the search process. Thus, we also attempt to extract the personal details discussed earlier. Unfortunately, pulling out geographical details from a person's profile can be burdensome. A user might list a hometown that is more specific than the FB geographic network that the town belongs to (i.e., San Jose is present in the Silicon Valley geographic network), which is problematic when assigning a worker to process the search query. Also, users typically do not list the country associated with a city. Since specifying a country during a search is less prohibitive, we prefer to search based on a country prior to including the city term. We used the Yahoo Location API to process any location-related information. Suppose a person lists that he is from Bullaburra, but never specifies what country this refers to. We issue an API call to Yahoo to resolve this location.

- **Extracting Email Addresses from Images:** Since hand labeling profile matches is quite laborious, we use matching profiles discovered via email addresses to augment our training set. Thus, we also include email addresses in our list of searchable attributes. Extracting emails is fairly easy when they are presented in text form; we simply create a regular expression matching the form of an email address. Facebook, however, displays email information in the form of images. To process these images, we used Tesseract [14], an OCR engine, that was trained on the Verdana font set. We achieved fairly accurate results from this tool, and we corrected our search results to the best of our abil-

ity, using known semantics regarding email addresses. For example, if Tesseract outputs a capitol letter I, we know this most likely corresponds to a 1, since all letters on Facebook email images are lower case.

- **Wrapping and Automating Search Calls:** When finding candidate matches, our workers utilize the search facilities provided by each OSN. However, these search facilities do not export an API to third-party programs like our workers. Thus, significant reverse engineering is needed to determine the appropriate HTTP requests corresponding to each search type. Facebook uses AJAX to standardize the names of educational institutions and companies. Since most scraping packages do not process Javascript, we manually inspected the HTTP headers to determine how to specify these fields.

  Finally, we have to deal with the possibility that the semantic information may be inconsistent across networks: for instance, a user may have moved, updated his new location on Facebook, but failed to do so on MySpace. Thus, using a simple conjunction of semantic fields as a match criterion may miss valid matches. To handle this, we progressively apply filters until we achieve the smallest number of search results using our search chains. For example, MySpace's search chain consists of name, gender, age, country and hometown.

- **OSN Idiosyncrasies:** OSNs vary considerably in potential geographic segmentation (e.g., Facebook is segmented by networks, MySpace is not) and the types of search queries they allow. We created 34 accounts on both OSNs in our study, registering each account with a specific geographic network. Because of the segmentation issue, we collect seeds around the geographic networks associated with the created accounts. While we would have liked to select completely random profiles, Facebook does not permit one to access profiles at random. Thus, we gathered our seeds randomly from the most populous cities/countries, where both FB and MS have significant presences. We focused primarily on English speaking countries, as we later train our classifier based on textual features and hence, we must be able to identify matching profiles purely by text.

- **Candidate Matching:** Given a seed profile S with a set of semantic fields (features) and a set of candidate matches for S, no obvious algorithm exists to determine whether a candidate is indeed the same user as S. We now discuss the method we employed using machine learning.

5.2 Matching Methodology

Once we have completed our seeding and searching, we are still faced with the task of determining which profiles in each OSN correspond to the same person. Recall that when we pick a seed profile S in OSN X, we extract searchable attributes from that profile and query OSN Y using those parameters. OSN Y will return a candidate set C, where $|C|$ may be greater than 1. We need some way of best determining which member of C actually corresponds to S, if any. If we found C based on an email address or a website reference listed in S, then no disambiguation is necessary. However, if we actually issued a search query, we need a more sophisticated method for profile matching. We use a technique from machine learning called boosting. Our features consists of various attributes that individuals list in their profile, including birthday, age, location, hometown, education, etc. Recall that during the mining process, we labeled these attributes in order to conduct our searches. For each feature, we do the following:

Let the bag of words from the seed profile for feature F_i be $S_{w,i}$, and let the corresponding bag of words from candidate profile P be $P_{w,i}$. We compute the intersection of the words $I_i = S_{w,i} \cap P_{w,i}$, as well as the union of the words $U_i = S_{w,i} \cup P_{w,i}$. We take the ratio I_i/U_i, and let that be our value for the feature. For example, if F_i is location, and $S_{w,i} = \{\text{San, Francisco}\}$, while $P_{w,i}$ is {San,Francisco, CA}, then $I_i = \{\text{San, Francisco}\}$, and $U_i = \{\text{San, Francisco, CA}\}$, giving us 2/3 as our feature value.

We added in several other features that are slightly more complicated. In particular, we used the friend name overlap count, which is computed by determining the number of friends with matching names between two profiles. We also added the number of search results returned during the querying process, since the higher the number of results returned from any OSN, the less confident we are that the individual is unique. We take the feature vector $F = (F_1, F_2, \ldots, F_n)$, and feed this into an open source boosting package called Jboost [7] with the default configuration parameters, setting the number of iterations to 200. JBoost outputs a classifier generated in python.

Boosting is a methodology to combine several weak hypotheses into a strong hypothesis. Let the labels be in $\{-1, 1\}$, then boosting creates a strong hypothesis $H(x)$ in the following manner:

$$\text{H(x)= sign}(\sum_{i=1}^{T} \alpha_i h_i(x))$$

where x is the feature vector, h_i is a weak hypothesis, and α_i is a weight assigned to each h_i. A weak hypothesis is a decision rule that exhibits weak predictive power towards the label. An example of a weak hypothesis is a so-called stump, which is a binary rule defined by thresholding a feature at a certain value. Formally, a stump h_i is:

$$h_i(\text{x}) = \begin{cases} 1 & \text{if } F_i \leq \theta_i \\ -1 & \text{otherwise} \end{cases}$$

where F_i is the feature under consideration, and θ_i is the threshold that separate the two classes of labels. Jboosts default configuration uses a conjunction of stumps as the weak hypotheses.

For both training and testing of our classifier, we require ground truth. We obtained a set of ground truth examples by generating 599 matching profiles via email addresses, and 286 via links to other OSNs. To avoid bias towards profiles with email addresses and links, we also *manually* labeled 200 additional profiles as matching and 500 as not matching. We conducted our manual verification by checking that user pictures and other markers (e.g., education) matched. We wrote a tool to make this manual verification process easier for a human.

The final ground truth set included 1385 matching pairs and 500 pairs of profiles that did not match. The feature vectors from 600 matching pairs (randomly selected 1385)

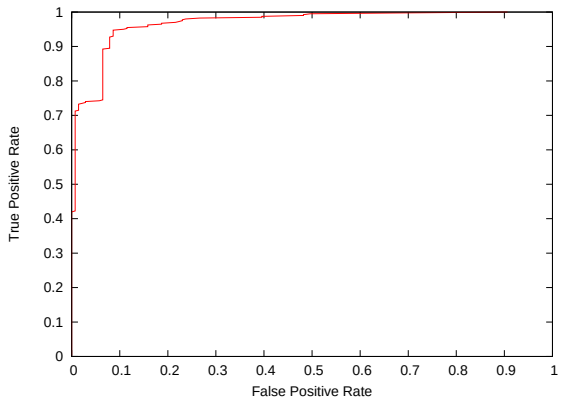

Figure 2: ROC curve showing the false positive rate versus the true positive rate for a test corpus.

and 300 non-matching pairs (randomly selected from 500) were used to train the classifier. Subsequently, 300 matching pairs and 200 non-matching pairs were used to test, as shown in Section 5.3.

5.3 Matching Validation and Threshold Selection

Our statistics are crucially dependent on the performance of the final matching classifier, especially when we lack an email address or a direct link to a profile on another OSN. We took a test set consisting of 300 matched and 200 unmatched profiles from the ground truth set described in the last section and ran the classifier against this set (these profiles were **not** used in the training data). The higher the value produced by the classifier, the more confident we are that the two profiles match.

The classifier produced by boosting computes a final numeric value that we must threshold using some threshold value T. Increasing T decreases the false positive rate but also decreases the true positive rate. By positive, we mean that a match output by the classifier agrees with the label present in the test corpus. Plotting the false positive rate against the true positive rate as T is varied yields the ROC (Receiver Operating Characteristic) curve shown in Figure 2. We set our threshold to be restrictive, resulting in a false positive rate of less than 5%, towards the left in Figure 2. We prefer to underestimate the amount of overlap. Our results can thus be considered a lower bound on the amount of overlap between OSNs.

6. ANALYSIS OF THE DATA

The data collection phase was conducted on May 12th, 2009 for 24 hours. We chose this duration to avoid consuming undue server resources for both Facebook or MySpace. The crawling phase generated 6654 seed profiles. We also extracted a subset of each seed's friends, producing an additional 12391 profiles. The search phase of our data collection yielded 68277 candidate profiles. We now describe our results for basic overlap, analysis of search attributes, active overlap, privacy conflict, and friend variation.

6.1 Basic Overlap

Using the classifier, we computed the basic overlap between Facebook and MySpace. We found that 25.2% of the FB profiles overlapped in MS, and 27.56% of the MS profiles overlapped in FB. When we compared our results with that of compete.com, we found that our results for the FB-to-MS overlap were closer to the estimate computed in Section 3. Our results for the MS-to-FB overlap are slightly higher (Compete measured a 20% overlap), but one may attribute this to an increased number of MS users joining FB.

Search Key Combination	Percentage of Matches
age,gender,location,name	1.78
age,location,name	0.36
country,gender,name	0.04
country,name	6.96
email	11.7
gender,location,name	0.06
href	5.7
location,name	12.5
name	37.82
name,school	19.2
name,work	3.88

Table 2: Percentage breakdown of which search key combinations led to matching profiles.

6.2 Search Attributes

Table 2 shows how each search attribute contributed to to the discovery of matching profiles across both OSNs. The name field (in isolation) contributed to slightly over 35% of our matches. This is not surprising, since the details listed on the same person's profile on two different OSNs may not strictly match. Combining the name field with a person's educational history accounted for a large portion of our matches (approximately 20%). Specifying a person's name and place of employment only contributed to 4% of our matches. Using a person's country, city, hometown or postal code in addition to his/her name produced 20% of our matches. Emails only contributed to 11% of our matches, since MySpace profiles do not typically list email addresses. The href row refers to the set of matches that were discovered by following a direct link from the seed profile.

6.3 Active Overlap

Of those individuals who maintain multiple profiles, we wish to look at the extent to which they remain active on every account they possess. Intuitively, while Section 6.1 describes what we might call formal overlap (a user is considered to be part of an OSN if the OSN registers the user as a member), we want to compute *active overlap* (a user is considered to be an active member of an OSN only if the user has shown some sign of activity on the OSN).

In order to quantitatively measure this, we mined 2055 overlapping profile pairs, where both pages in the overlap contained some type of date indicative of activity. We extracted the dates by handcrafting parsers for each type of downloaded page (e.g., MS Comment Page, FB Blog Post Page). For example, Facebook records the dates on a user's Wall Page when that individual performs some type of action, such as commenting or posting a link. Using Hpricot, we wrote XPath expressions to extract these dates. We consider two matching profiles as actively overlapping with parameter T if the difference between the last activity times on the two matching profiles is less than T. Clearly, as T goes to infinity, active and basic overlap will become identical. A striking observation from Figure 3 is that only 58%

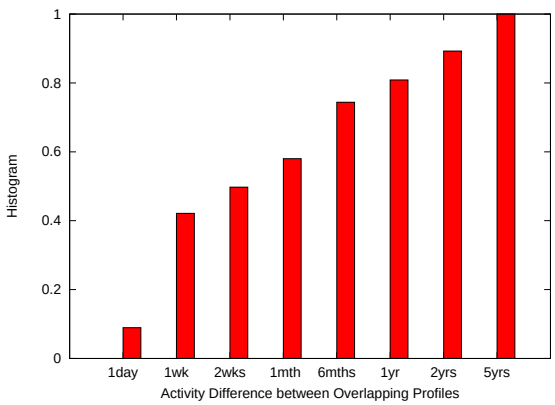

Figure 3: Measuring overlap for various values of the recentness parameter T ranging from 1 day to 5 years.

of users who are members of these networks show activity concurrently within a month. This indicates that users do not actively monitor their various online identities.

Note that our basic overlap numbers of around 25% provide a good prima facie case for OSN aggregators [9]. On the surface, a quarter of the members of MySpace and Facebook are members of another social network and can benefit from software that synchronizes their profiles on both networks. However, the active overlap numbers undermine this argument. This is because a majority of users that belong to both Facebook and MySpace apparently do not mind letting one of their two profiles remain out of date for a month.

6.4 Privacy

While crawling FB and MS, we noted that users take different approaches to privacy depending on what social network they belong to. We looked at 5296 overlapping profiles, and determined what type of security policy the user specified (i.e., public or private). In 2319 of the overlapping cases, the profiles agreed on the privacy settings, meaning that the profiles were observable from both the OSNs crawled. In the remaining 2977 cases, the user specified conflicting viewing policy interests, leaving one open while restricting access to the other. In general, we found that Facebook users tend to leave their profiles open to individuals in the same geographic network, while the same individuals tended to have restricted MySpace profiles. This is most likely due to the copious number of spammers present on MySpace [16].

6.5 Friend Variation

The active overlap numbers suggest that one explanation why users belong to more than one OSN is due to historical reasons. At some point in time, the OSN was popular, prompting the user to join. However, the user subsequently became inactive in that OSN. However, a second reason for multiple OSN memberships is that a user belonging to FB and MS may have disjoint friends in those networks, neither of which belong to the other network. In that case, the user wants to stay in touch with more friends by joining multiple OSNs. We look at the absolute difference in the number of friends an overlapping user has in both FB and MS.

If a user has 200 more friends in Facebook than in MySpace, the user must have 200 *different* friends in both networks. However, if a user has 50 friends in both Facebook and MySpace, the user may or may not have 50 different friends in both networks, which the Friend Overlap measure attempts to estimate by sampling and matching.

avg	std	med	total profiles
210.724	232.452	138.0	4418

Table 3: Friend count differences between overlapping profiles

We looked at the friend variation count for 4418 profile pairs, where both profile listings contained a visible friend list. We computed the average, standard deviation, and median of the absolute value of their friend count differences. Note how the overlapping profiles show a huge variability in their friend counts (an average difference of around 200 friends), meaning that perhaps users are participating in different OSNs to achieve greater reachability.

7. RELATED WORK

The measurement of OSNs is emerging as a field of study that employs many techniques previously used to measure computer networks. In this vein, [1] and [11] perform an analysis of the graph structure of social networks, motivated by earlier studies of the Web and Internet graphs. The results consisted mostly of graph measurements, such as the assortativity and verifications of power-law and scale-free properties of such graphs.

Our paper differs from these in methodology (analysis of content versus graph structure) and results (graph versus overlap analysis). [8], which shows that geographical proximity is related to friendship, is closest to our paper in spirit, as the question requires some content analysis content. However, our motivation is different and requires more sophisticated content analysis, of which location is a small part.

[2] focuses on identity theft, but presents an algorithm for matching individuals across social networks to conduct cross-site profile hijacking. The algorithm assigns points based on whether the name and education fields match between two profiles. The authors do not thoroughly evaluate the efficacy of their techniques, since the main focus of the paper is not to locate people. Several algorithms for matching duplicate entities in databases (i.e., record linkage/entity resolution) are described in the survey [4]. Several of these techniques (e.g., token-based edit distance) may improve the accuracy of our classifier; however, the primary focus of this paper is investigating various combinations of search keys, for which the classifier is only an evaluation mechanism.

At least two high-profile estimates of basic overlap have been done in the commercial sector. In 2007, compete.com [3] published a matrix of OSN overlap. Compete obtained its data by instrumenting a random sample of users with measurement code that monitors their web accesses. In 2007, the online reputation aggregator Rapleaf [12] announced some limited overlap results based on user provided information. While these studies are valuable, they are not repeatable, and the companies do not provide access to their data. Further, they do not mine the data for finer-grain information (e.g., friend overlap, recentness) that provide insight into why social networks overlap.

Finally, [5] reports on a survey administered to 1060 first-year students at the University of Chicago, Illinois. While there is some rough tabulation of overlap, the objective of that study was to explain social network usage with respect to other variables such as parental education. Further, the

sample's bias towards college students may have resulted in fairly large overlap numbers for MySpace and Facebook.

8. CONCLUSIONS

This paper discusses a system for searching and matching individuals on Facebook and MySpace. Our system allows us to present a quantitative study of user overlap in both OSNs. Unlike earlier measurement studies of OSNs, our searching and matching methodologies focus on analyzing content to obtain features that are essential for matching, such as country and email addresses. Furthermore, we believe our resulting analysis of OSN overlap provides insight into the dynamics and structures of social networks. We also used tools from learning theory and natural language processing that can assist in the analysis of OSNs. We arrived at the following conclusions in this paper:

- Based on our study, we observe that specifying certain attributes during a search increases the likelihood of locating a person. For example, using an individual's name and educational institution contributed to five times the number of matches we discovered compared to using the person's name and workplace. We show that using search chains of increasing specificity may be a better search interface than employing a simple conjunction of attributes (as is offered currently by FB and MS). Currently, users must try many different conjunctions of attributes during a search; search chains attempt to alleviate this problem.

- Our study concludes that 27% of MySpace users also have accounts on Facebook. This agrees with earlier numbers reported by compete.com. However, our study reports that around 25% of Facebook users also have accounts on MySpace, compared to 64% for compete.com. In the last two years, Facebook has more than doubled, while the size of MySpace has remained roughly stagnant. Thus, only a small fraction of new Facebook users may have possibly joined MySpace.

- 58% of users that are members of both Facebook and MySpace have not been active in more than one of these networks during the last month. This undermines the case for social aggregation software.

- When users belong to both Facebook and MySpace, their friends rarely overlap. More precisely, there is an average difference of at least 200 friends. This lends credence to the hypothesis that users join both Facebook and MySpace to stay in touch with different sets of friends.

For future work, we hope to build on our architecture and incorporate image-based techniques during our matching phase. With the rise of FB, we imagine that more people will migrate away from their previous OSNs; we envision building a tool that will enable this migration. Furthermore, we hope to investigate other questions about online identity maintenance. For example, when a user is a member of two OSNs, we wish to quantify whether there are differences in the content the user puts in his or her profile in the two OSNs. Measuring content difference is non-trivial, but we believe that such measurement can lead to valuable insights. More generally, we believe that content-analysis can lead to a host of questions that can be applied to better understand the structure, form, and evolution of social networks.

9. REFERENCES

[1] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM.

[2] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All your contacts are belong to us: Automated identity theft attacks on social networks. In *18th International World Wide Web Conference (WWW2009)*, April 2009.

[3] Compete. `http://blog.compete.com/2007/11/12/connecting-the-social-graph-member-overlap-at-opensocial-and-facebook`.

[4] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions On Knowledge and Data Engineering*, 19, January 2007.

[5] E. Hargittai. Whose space? differences among users and non-users of social networks, journal of computer mediated communication, vol 3, issue 1, 2007.

[6] R. Holzer, B. Malin, and L. Sweeney. Email Alias Detection Using Social Network Analysis. In *Proc. LinkKDD*, 2005.

[7] JBoost. `http://jboost.sourceforge.net/`.

[8] D. Liben-Nowell and et al. Geographic routing in social networks. In *Proceedings of the National Academy of Sciences (PNAS)*, pages 11623–11628, 2005.

[9] Mashable. `http://mashable.com/2007/07/17/social-network-aggregators/`.

[10] M. McNett, D. Gupta, A. Vahdat, and G. M. Voelker. Usher: An Extensible Framework for Managing Clusters of Virtual Machines. In *Proceedings of Large Installation System Administration Conference*, 2007.

[11] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, New York, NY, USA, 2007. ACM.

[12] Rapleaf. `http://www.nexdot.net/blog/2007/11/12/social-network-overlap-and-why-opensocial-could-be-useful/`.

[13] F. P. Room. `http://www.facebook.com/press/info.php?statistics`.

[14] Tesseract. `http://code.google.com/p/tesseract-ocr/`.

[15] VentureBeat. `http://venturebeat.com/2009/01/07/facebooks-traffic-growth-leaving-rivals-in-the-dust/`.

[16] S. Webb, J. Caverlee, and C. Pu. Social Honeypots: Making Friends With A Spammer Near You. In *Proceedings of the Conference on Email and Anti-Spam*, 2008.