An Unsupervised Approach to Discovering and Disambiguating Social Media Profiles

Carlton T. Northern
Old Dominion University
Department of Computer Science
Norfolk,VA 23529 USA
carlton.northern@gmail.com

Michael L. Nelson
Old Dominion University
Department of Computer Science
Norfolk,VA 23529 USA
mln@cs.odu.edu

ABSTRACT

Social media in the last decade has become a popular communication mechanism on the web. Sites like Facebook, Twitter and YouTube are seeing enormous growth. It is important to understand the trends of this new type of media for many reasons including identity theft, social engineering, advertising and digital preservation. Some data sets have been made available to the public such as the tweets from Twitter, alternately data can be scraped from the open web. However, to ascertain trends from a group of individuals such as employees of a business, or students of a university, there is no way, without asking each individual member, what social media sites they use. Within this paper, we present a detailed approach to gaining this type of information. Specifically, for a group of geographically and organizationally affiliated members, we present an unsupervised approach that can discover and disambiguate social media profiles with a precision of 0.863 and an F-measure of 0.654.

Categories and Subject Descriptors

H.5.4 [**Hypertext/Hypermedia**]: User issues

General Terms

Algorithms, Experimentation, Measurement

Keywords

Web Science, Social Media, Identity Disambiguation, Entity Resolution, Semantic Web

1. INTRODUCTION

Over the last decade, social media has become a popular communication mechanism on the web. Sites like Facebook, Twitter and YouTube are seeing enormous growth. It is important to understand the trends of this new type of media for many reasons including identity theft, social engineering, advertising and digital preservation. It is also important to understand trends of this media across sites, and not just the trends of individual sites. For instance, is there more personally identifiable information for users of Blogger than for users of Twitter? However, our initial interest in studying the use of social media sites was to try to discover explicit or implicit digital preservation strategies present in the user population of a computer science department. In the past, the students' `public_html` directory on the departmental server was the best way to host files and claim an online identity, but `public_html` is all but extinct for our computer science students (we only found 316 out of 2016 accounts that had `public_html` directories, of those only 140 contained `index.html` (or similar) files, and of those only 53 contained the students name, indicating that it is being used as a profile). Upon reflection, it was obvious the student population had migrated their files and identities to social media sites.

Some data sets of social media have been made available to the public for download such as a recent release of tweets from Twitter[1]. Alternately, data can be scraped from the open web much like how Google indexes the web. However, to ascertain trends from a large group of individuals such as employees of a business, or students of a university, one must first know which social media sites these individuals are using. To do so, one must check each social media site of interest for membership of each individual. For a large group, manually performing this work is out of the question. So, we look for an automated solution that can find these profiles and present the findings with the most amount of accurate results and the least amount of profiles not belonging to the individual of interest. In other words, precision is more important than recall because its more important to perform research on nothing at all rather than do it on the wrong person.

Using search engines to find this type of data will produce good recall, but very low precision. For instance, searching for the person "Michael Nelson" on Google will produce 274,000 results. Obviously not all the resources are actually referring to the Old Dominion professor Michael Nelson, and a decision needs to be made whether or not each resource refers to him, or a namesake (multiple people that share the same name). This decision is referred to as disambiguation. If we modify our query to include location and school affiliation "Norfolk" and "Old Dominion" we can reduce the results to 9,800. Still, much more filtering of results is needed to

[1] `http://www.readwriteweb.com/archives/twitter_data_dump_infochimp_puts_1b_connections_up.php`

achieve the real-life 7 profiles that exist for the Michael Nelson. To make matters worse, some social media sites, such as Facebook have user-defined settings to opt-out of search engine indexes making these profiles unaccessible from search engine queries.

One could alternately poll each social media site of interest for an individual by dereferencing profile pages with a given username. This approach is used by `Knowem.com` which checks about 300 different social media sites for usage of a username. However, this supposes that the username of the individual for the site is known. A similar method can be performed with email addresses. By signing up for an account on a site with an individuals email address, one can check for an error message stating that the email address is already in use, if so, the user has an account. This technique is used by another service called `Rapportive.com` which is a CRM tool that sits on top of Gmail. Once again the individuals primary email address(es) must be known ahead of time.

The approach used in this paper uses a blending of these approaches to discover possible profiles. It then uses a set of heuristics to disambiguate the profiles we are interested in from those that we are not. The heuristics used are a combination of keyword matching, community structure analysis, and extraction of semantic and feature data from profiles. A simple scoring system is used to track the presence and absence of certain features and when a candidate profile passes a certain point threshold (11 of approximately 50), the profile is validated and included in our graph analysis.

We have tested this approach on the Unix login ids of the computer science students of Old Dominion University. From the department, 2016 ids were selected with only their "finger" information (full name, and `cs.odu.edu` email) known:

```
% finger mln
Login: mln              Name: michael nelson
Directory: /Users/mln             Shell: /bin/tcsh
New mail received Sat Apr 31 23:19 2011 (EDT)
     Unread since Sat Apr 31 23:18 2011 (EDT)
No Plan.
```

Evaluating our results requires the real set of social media profiles for each student, which was unattainable, so a sample set of 22 members from our internal research group were selected to provide a truth set of their social media profiles. We used this collection of 22 people and 139 self-reported social media profiles from our truth set, from which our automated scoring approach yielded 0.863 precision, 0.526 recall and an F-measure of 0.654.

We have structured the paper as follows: Section 2 presents related research. Section 3 details the requirements of the approach. Section 4 presents the proposed approach to discover and disambiguate social media profiles on the web. Section 5 presents the results of an experiment conducted that tests and measures our approach. Section 6 presents the conclusions drawn from our work.

2. RELATED WORKS

Approaches designed to find and disambiguate social media profiles for a large group of individuals belonging to the same organization are not very common. However, it can be looked at as a subset of people search. An unsupervised monitoring approach described in [10] extracts information from web pages such as lexical, linguistic and personal information. They then apply a clustering strategy to a set of gathered web pages, resulting in clusters of pages for given person. One major limitation of unsupervised clustering approaches is their unfocused nature and the need to discover how many different namesakes are referred to in the set of resources.

Rowe and Ciravegna [6] presented a semi-supervised machine learning approach for disambiguating identity web references (i.e. web resources containing information about a person) where seed data is exported from a social network and used to train their algorithm. They achieved a very high precision level and high F-measure, but the requirement for seed data is one that cannot always be fulfilled.

This work is also closely related to Entity Resolution (ER) research done in the context of digital libraries, including techniques mining authorship graphs [4, 2], using search engine results to find overlapping patterns of co-authors [7], and clustering the results of various ER subsystems [3].

We believe our approach is unique in combining a simple scoring system, tailored to social media sites, not requiring highly structured input (e.g., co-authorship graphs), and little to no supervision.

3. REQUIREMENTS

To find and disambiguate social media profiles for a large group of individuals, we imposed the following criteria for a feasible approach.

1. The approach must be completely automated with the only human interaction being the creation of a search query consisting of a location, an organization, and a profession/education domain.
2. Achieve a precision of 0.85 or higher. It is important to achieve a high precision because it would be better to perform no research at all than to perform research on the wrong profiles. A value of 1.0 is unnatainable because the impact to recall would be devastating.
3. Achieve a recall of 0.5 or higher which is comparable to a human level of processing with a precision of 0.85 [5]. With precision and recall fixed, this results in a needed F-measure of approximately 0.63.
4. The approach must be able to find profiles which are not indexed by current search engines such as Google, Yahoo, MSN.
5. The approach can use any publicly available web services, such as using search engine, semantic search engines, page scraping, and well-known or undocumented web APIs.
6. Only publicly declared identities are of interest; we do not seek to map user names to obfuscated identities (e.g., "Bruce Wayne" → "Batman").

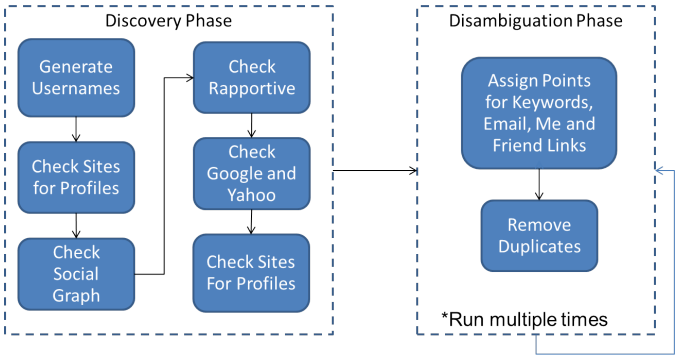

Figure 1: Algorithm

7. The approach must find profiles from 25 pre-defined sites selected first for (our assessment of) their popularity and secondly to provide a balanced mix between social networking, photo/video sharing, blogging/microblogging, social music, community of interest based, and social news sites. Some of the more notable sites include Facebook, LinkedIn, Blogger, and Flickr. A complete list of these sites can be found in Table 4.

8. New social media sites can be added to the list as appropriate and with minimal change to the baseline code.

4. APPROACH

Our approach is divided into two main phases, *discovery* and *disambiguation*, depicted in Figure 1. The discovery phase's purpose is to seek out social media profiles for an individual by utilizing a combination of search engine queries, semantic web queries and by polling social media sites themselves for existence of profiles with probabilistic names of the individual. The disambiguation phase's purpose is to then whittle down the results found in the discovery phase into only those that are the individuals we are concentrating on, which we will call positive results. The heuristics used in this phase are a combination of keyword matching, community structure analysis, and extraction of semantic and feature data from profiles which culminates in a point based value that determines if a profile is positive.

4.1 Discovery Phase Approach

The discovery portion of this approach starts with the polling of each social media site to check if a profile exists for a username that is a variation of the individuals name. The usernames created from the individuals name, using an example, are: [michaelnelson, michael.nelson, michael_nelson, michael-nelson, mnelson, nelsonm].

If there is a known username for an individual, such as a university or work account username (e.g., "mln"), it is also checked at this point. Then, for each of our 25 social media sites, an HTTP GET request is made to the site for the page that holds a users profile. For instance, www.facebook.com/michaelnelson is retrieved, and if it returns a 200 response the profile and it's content is saved. If it produces a 404, or soft 404, (i.e., meaning that a 200 response is sent but a "not found" error message is displayed in the page itself, see [1]), it is discarded. Soft 404's are somewhat problematic from a code maintenance perspective because it forces the code to take into account whether or not an error message is displayed on the page, and these error messages can change, meaning that code will break over time.

4.1.1 *Google's Social Graph API*

The profiles that were previously found are then used as queries for Google's Social Graph API[2]. The Social Graph API operates over the entire web looking for hCard + XHTML Friends Network (XFN) microformat data[3] and Friend of a Friend (FOAF)[4] RDF data.

By querying the Social Graph API for each profile and looking for "me" links (links where a rel value is set as "me" to infer ownership), we find other profiles that we may not have otherwise been exposed to by heuristic name variations or search engines. Another benefit that the Social Graph API provides is that new usernames can be extrapolated which can then be used to cross reference our social media sites. Indeed, these usernames are saved for later use.

4.1.2 *Rapportive*

Rapportive is a social contacts relationship management (CRM) tool that plugs into Gmail. It's sole purpose is to provide information about the person for the currently read email. It does this without any coordination/cooperation from the individuals that it profiles. Effectively, Rapportive is working on the same problem that the approach of this paper is trying to solve, that of finding social media profiles. Rapportive is somewhat limited in the seed data they have for an individual only having access to a users name and an email address. Theoretically, Rapportive could mine the contents of emails and use this data for an unsupervised clustering approach, but it is unclear at the time of this writing to what extent Rapportive uses the information that is available to them. Figure 2 is a screenshot for Michael Nelson's Rapportive results.

Rapportive does not have a public API, however, they operate using a RESTful web service that returns JSON results. The only authentication needed is an initial handshake when the user is logged into Gmail. By inspecting the calls made from the browser one can ascertain which call is being made to Rapportive and then replicate this call on their own with any email address used as the query. By using this method, calls to Rapportive with an individuals email address will yield all the results that Rapportive is holding for that individual. At the time of this writing, results are not very comprehensive. It seems that Rapportive, like the requirements of our approach, favor a high precision rate rather than high recall. However, Rapportive did not contribute significantly to to our experiment as we shall see in the next section.

Once results have come back from both the Social Graph API and Rapportive, the newly discovered usernames from these services are ran through the polling process again.

[2] `http://code.google.com/apis/socialgraph/`
[3] `http://gmpg.org/xfn/`
[4] `http://www.foaf-project.org/`

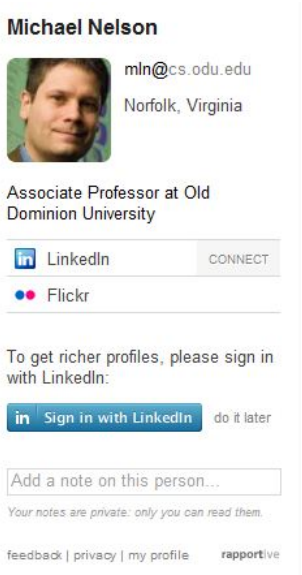

Figure 2: What a Rapportive User Sees When Reading an Email From Michael Nelson

This achieves the effect of finding usage of other usernames that we had not previously generated.

4.1.3 Google and Yahoo Search

The next step of the discovery phase is to query the Google Search API and the Yahoo Boss Search API. Both Google and Yahoo limit the number of results provided through their services and so it is important to make the queries count. Simply querying for the individuals name is not sufficient because it is too broad of a query. To make the queries more specific, the individual's name is used in addition with either a location, a profession/education domain, or a specific site. The following queries are representative of the type of queries used:

1. "michael nelson" AND norfolk
2. "michael nelson" AND "computer science"
3. "michael nelson" AND "old dominion"
4. "michael nelson" site:http://www.facebook.com

To generate these search terms automatically we can look to the semantic web. First of all, surrounding city names can be derived from GeoNames[5] starting with only one location/city. Terms related to profession or education background can be derived from WordNet [8]. Lastly the organization's name is known and the social media sites of interest are known.

Different result sizes were tried from 8 to 64 and results were recorded and compared against a truth set of 18 accounts. In every case except for one the matched profiles were in the first 1 - 8 results. This drastically reduces the amount of profiles to disambiguate in the next phase by cutting out the results from 9 - 64. Queries past 40 were not performed for Google because API limits. Table 1 contains these results.

[5]`http://www.geonames.org`

Table 1: Search Rank Results

	Positive		Negative	
	Google	Yahoo	Google	Yahoo
1 - 8	13	22	325	286
9 - 16	0	1	277	205
17 - 24	0	0	244	125
25 - 32	0	0	206	73
33 - 40	0	0	193	71
41 - 48	-	0	-	73
49 - 56	-	0	-	14
57 - 64	-	0	-	0

4.2 Disambiguation Phase

The disambiguation phase's purpose is to validate the candidate results found in the discovery phase. The heuristics used in this phase are a combination of keyword matching, community structure analysis, and extraction of semantic and feature data from profiles: each match on one of these features yields "points", varying from 1 point for weak indicators (e.g., keywords like "programmer") to 10 points for `rel='me'` links. Similarly, the absence of a person's name (and associated variations as described in 4.2.1) results in a deduction of 21 points; the heavy penalty reflects our preference for precision over recall. As described below, candidate profiles can score as high as 50+ points (the number can vary depending on the presence of keywords), but we set a threshold of 11 points for a candidate to be considered validated. This threshold was chosen because its the value in which many combinations of points predict a true positive profile. For instance, the existence of a `rel='me'` and a keyword, or a full name and a community structure link, both equal 11.

4.2.1 Name disambiguation

Names play a very important role in the disambiguation process. When searching over a large group of individuals where nicknames, diminutive and even middle names or middle initials are unknown. To make matters worse, some sites, like Twitter, encourage the use of aliases rather than real names. Rules and strategies must be created to handle these edge cases. One such strategy is to use a database of diminutive and nicknames to so that you can then find individuals that shorten their name (e.g.. "Jefferey" shortened to "Jeff"), or use nicknames (e.g., "Robert" nicknamed "Bob"). Since we were unable to find a pre-existing database or service (we could only find expensive commercial products), we created our own solution pieced together from a genealogy naming web site[6]. Note that our mapping of nicknames/diminutive names to formal names is only for English language names.

With nicknames and diminutive names in place, profiles are keyword searched for first, last, diminutive, and nicknames. If a first name is found the profile is assigned 2 points, if the

[6]Available at: `http://code.google.com/p/nickname-and-diminutive-names-lookup/`

last name is found, it is assigned 4 points, if a diminutive or nickname is found it is assigned 2 points. This results in a total point value of 6 if a last name is found and either a first name or a diminutive or nickname is found.

This procedure is performed for every social media site except for SlideShare, Facebook, Blogger, LinkedIn, and Google Profiles. These sites are treated separately because they tend to promote the use of the user's actual real name rather than an alias (note that this isn't always the case because the user can choose to use an alias). For these sites, the name is derived from the HTML and then either awarded points for a name, or taken points away for a non-matching name. If a first, diminutive or nickname is found for these sites, it is awarded, 2 points. If a last name is also found then it is awarded another 5 points. However, if both of these are not found, then 21 points are subtracted because we know that the names do not match and it likely belongs to someone else. This results in high fidelity name matching for these 5 sites.

4.2.2 *Keyword Disambiguation*

Keyword disambiguation is the same as the name disambiguation using keywords except for rather than using names, it is using the location or profession/education domain keywords that are also used in the Google and Yahoo searches. 7 points are awarded for keywords that match nearby cities. 9 points are awarded for instances where the keyword matches the individuals organization. 4 points are awarded for profession/education domain phrases that are two word based, such as "computer science", and 1 point is assigned for one word, such as, "programmer".

4.2.3 *Me Links and Emails Disambiguation*

In the case that a profile is found from Rapportive from an email address 10 points are assessed to the profile. 10 points are also assessed to profiles that were found to be a "me" link from another profile that has already been evaluated to be a positive result. This action has the assumption though that every profile has already been disambiguated and so we must assess "me" links multiple times.

4.2.4 *Community Structure Disambiguation*

By operating over a set of people from the same organization and by the social nature of the content we are interested in, community structure present in hyperlinks can be used to aid the disambiguation process. For each page, hyperlinks are harvested and then evaluated to see if the page that is being pointed to is already present in the set. If it is a positive result, 5 points are awarded to the page containing the link. Figure 3 depicts this process. If it is not a positive result, but the page containing the link is a positive result, 5 points are awarded to the page in the link. No more than 5 points are awarded to a profile for this feature because while the presence of community structure is a good indicator that the individual at least knows individuals from the organization of interest, it does not indicate that it is indeed the individual of interest.

5. EXPERIMENT

5.1 Evaluation Metrics

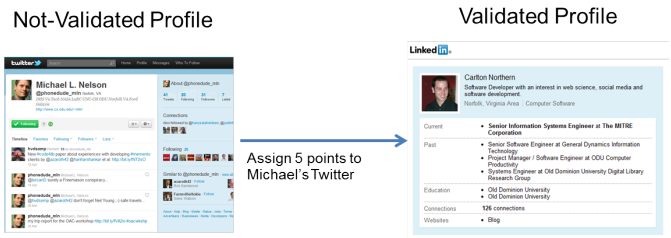

Figure 3: Community Structure Process

To assess our presented approach's ability to discover and disambiguate social media profiles we use the information retrieval metrics; recall, precision and F-measure [9]. A denotes the set of relevant social media profiles and B denotes the set of retrieved social media profiles, therefore:

$$precision = \frac{|A \cap B|}{|A|} \tag{1}$$

$$recall = \frac{|A \cap B|}{|B|} \tag{2}$$

F-measure provides the harmonic mean of both precision and recall. Let S be the set of all students $\{s_1, s_2, ..., s_n\}$. Then let:

$$precision(s_i) = \frac{|A_i \cap B_i|}{|A_i|} \tag{3}$$

$$recall(s_i) = \frac{|A_i \cap B_i|}{|B_i|} \tag{4}$$

$$F - measure = \frac{2 \times precision(s_i) \times recall(s_i)}{precision(s_i) + recall(s_i)} \tag{5}$$

5.2 Dataset

The dataset was copiled using the student body of the Computer Science department at Old Dominion University, sampled from February 2011. This dataset consists of the accounts of 2014 students and 2 professors. Of the 2014 students, 140 are graduates and 1874 are undergraduates. For each account, the first and last name, and CS email address was given from the Unix finger command.

5.3 Evaluation

Evaluating our results would require the knowing all the social media profiles of our dataset. This data was unattainable, so a truth set of 22 members and recent alumni of the author's research group Web Science and Digital Libraries (WS-DL) of Old Dominion University were selected and their real profiles recorded and used to measure the approach. These members vary from a high level of social media presence to low levels, from a high of 21 to a low of 1 profile (Table 2 summarizes the truth set results). The mean is 4.94 with a standard deviation of 4.14. Facebook was the

Table 2: Results From the Truth Set

Name	Positive Profiles	Known Profiles	Precision	Recall
C. Northern	13	19	1.0	0.684
M. Nelson	7	7	0.571	0.571
wsdl03	7	11	1.0	0.636
wsdl04	3	3	1.0	1.0
wsdl05	2	4	0.5	0.25
wsdl06	4	8	1.0	0.5
wsdl07	2	7	1.0	0.286
wsdl08	3	6	1.0	0.5
wsdl09	4	9	1.0	0.444
wsdl10	3	4	0.667	0.5
wsdl11	3	6	1.0	0.5
wsdl12	4	5	0.25	0.2
wsdl13	4	10	1.0	0.4
wsdl14	4	5	1.0	0.8
wsdl15	2	2	1.0	1.0
wsdl16	1	2	1.0	0.5
wsdl17	1	3	1.0	0.333
wsdl18	1	1	1.0	1.0
wsdl19	3	10	1.0	0.3
wsdl20	1	5	0.0	0.0
wsdl21	2	3	1.0	0.667
wsdl22	2	4	1.0	0.5
Total	76	139	-	-
Mean	3.454	6.091	0.863	0.526
SD	2.703	4.023	0.284	0.262

most used site of the truth set with 15 profiles, followed by Blogger and LinkedIn both with 12 profiles.

Although our intuition is that our truth set is representative of the data set, we cannot prove this. The members of our research group are either professors (2) or graduate students (20) and the mean age is undoubtedly higher than the undergraduate majority of the test data set. On the other hand, the nature of our research group probably increases awareness of social media sites and their applications. Given the difficulty of gathering accurate data and our desire to track the practices of the departmental community at large, we assume the truth data is representative.

5.4 Results

For our truth set, the approach achieved an average precision of 0.863, recall of 0.526 and an F-measure of 0.654. Recall was a limiting factor for our F-measure value; we had an intentional bias toward precision over recall. We believe the results are similar for our data set. The exact number of profiles for the truth data set for each social media service is presented in Table 4. Note that the profiles that passed validation include false positives. For example, for Michael Nelson there were seven truth profiles and seven positive profiles in our result set. However, only four of the seven positives were true positives (three were false positives), so the precision of 0.571 and recall 0.571 reflect this.

It should be noted that after the discovery and before the disambiguation phase, the precision was 0.064 and recall was 0.718. The very low precision is expected due to the

Table 3: Positive Social Media Profiles From the Whole Dataset

Service	Truth Set	Data Set
facebook.com	16	595
twitter.com	10	333
myspace.com	2	201
linkedin.com	13	180
profile.google.com	4	65
slideshare.net	4	58
blogger.com	14	56
stumbleupon.com	1	56
youtube.com	1	39
flickr.com	1	37
picasaweb.google.com	1	33
last.fm	0	22
delicious.com	4	15
tumblr.com	0	14
identi.ca	0	8
pandora.com	1	8
friendfeed.com	1	7
digg.com	1	4
reddit.com	1	4
newsvine.com	0	2
spock.com	1	1
technorati.com	0	1
eventful.com	0	0
mixx.com	0	0
tribe.net	0	0
Total	76	1739

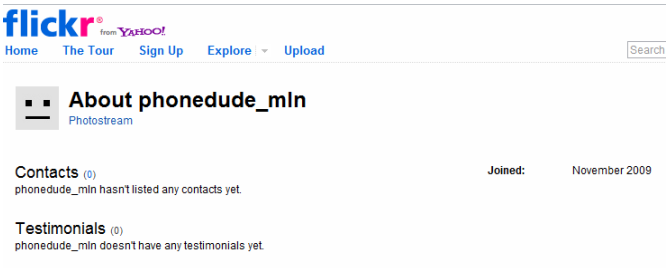

Figure 4: Flickr Has Little Identifying Information

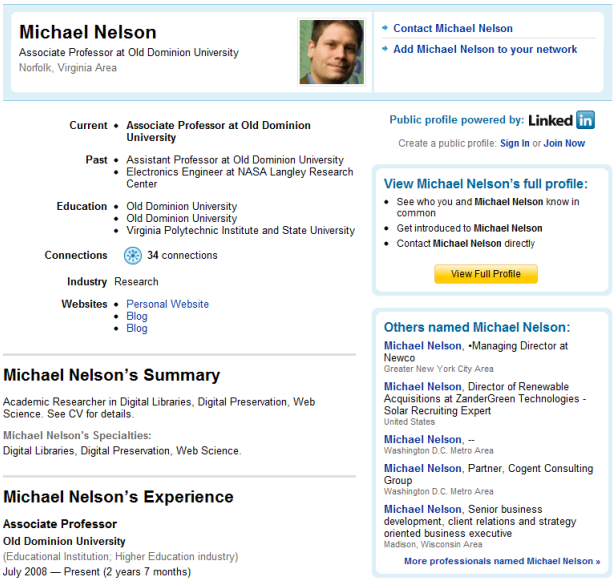

Figure 5: LinkedIn Is Rich With Identifying Information

high number of profiles (an average of 99 per person) found for an individual. So, the recall value is dropping 0.183 after disambiguation. This is largely due to profiles that are sparsely populated with personally identifiable information, or in some cases little to no content at all. YouTube and Flickr are common examples; figure 4 shows a Flickr profile that contains no identifying information. This profile was signed up for and forgotten about, it was not even reported in the truth set until our approach found it and it was realized that it should be included in the truth set. This happened multiple times with multiple people in our research group, and illustrates the challenge of establishing a test set.

In contrast, some social media sites tend to have a wealth of information, like LinkedIn. Figure 5 shows Michael Nelson's LinkedIn profile which contains his location, current and past employment including job title, education history, and it even contains "me" links to his other web sites.

One side effect of the community structure disambiguation is that it can sometimes produce false positives for profiles that do not belong to the relevant individual, but rather to one of their friends. Take for example a link from Dr. Nelson's Picasa profile to Carlton Northern's Picasa profile. This link contains Carlton Norther's name as the anchor. When disambiguating this profile in the set of Carlton Norther's links, it may be assigned a higher point value than Carlton Northern's actual Picasa profile because it will get an additional 5 points for the presence of a community structure link. This side effect is mitigated on Facebook, LinkedIn, Blogger, SlideShare, and Google, where the approach extracts an individuals name directly from the HTML in the profile.

Rapportive had provided fewer results than originally expected with only 15.9% of our truth set profiles being found from this source, and only 1.6% being unique to Rapportive and not found through other means. This is most likely due to the email addresses in which we had access to for this experiment, which are the ODU CS Department email addresses (i.e., `xxx@cs.odu.edu`). These email addresses are typically used by the students for school use only and are not necessarily used outside of school, including registering accounts on social media sites. In an attempt to gain access to the main email address that these students use, `.forward` files were examined on the ODU CS student UNIX accounts. A `.forward` file placed in a user's home directory will forward email to whatever email address(es) is/are in that file. However, only 25 alternate email addresses were found for the 2016 accounts from our data set. Many users had their `.forward` file set with permissions that were not world readable and thus we did not use this information.

Using a clustering approach as in [10] may also improve upon our recall score. This will be explored in future research. However, as discussed in [6] unsupervised clustering approaches as in [10] suffer from the need to discover how many different namesakes are referred to in the set of resources.

5.4.1 Graphs

To visualize our results, graph representations have been constructed. Figure 6 shows the the social media sites in use by our research group as nodes with varying sizes depending on how many positive profiles for that site are found in our truth set. The edges depict community structure and are the links to other positive profiles found in our truth set. Edge sizes are also based on the number of links found from the linking site to the linked site. The three most popular sites based on PageRank are LinkedIn, Twitter and Blogger. There are 15 weakly connected components in the graph and 25 strongly connected components. The average clustering coefficient is 0.041. The average degree is 0.96 with an average in degree and out degree of .48. The average path length is 1.0 and the network diameter is 1.

Figure 7 is the same as Figure 6 but over the whole 2011 data set. The three most popular sites based on PageRank are Facebook, LinkedIn, and Twitter. There are 11 weakly connected components in the graph and 21 strongly connected components. The average clustering coefficient is 0.147. The average degree is 2.56 with both an average in degree and out degree of 1.28. The average path length is 1.692 and the network diameter is 4.

In Figure 8, the nodes are students, weighted by the amount of positive profiles and the edges are links between them via any of the social media services (e.g., Carlton's blogger.com profile links to Michael's YouTube profile). Our research group is colored red and graduate students are colored green (our research group is comprised solely of graduates). We omit the graph for the truth set since these results are highly clustered and can be viewed by zooming in on this graph. It was the authors assumption that the graph would be more densely populated with edges, but upon inspection this shouldn't be the case. If this graphs were created with *all* links between students known (i.e. every Facebook friend, every Twitter friend, every LinkedIn connection, etc.) the graph would be densely populated. However, we are showing only the links found on profile pages, which upon a cursory

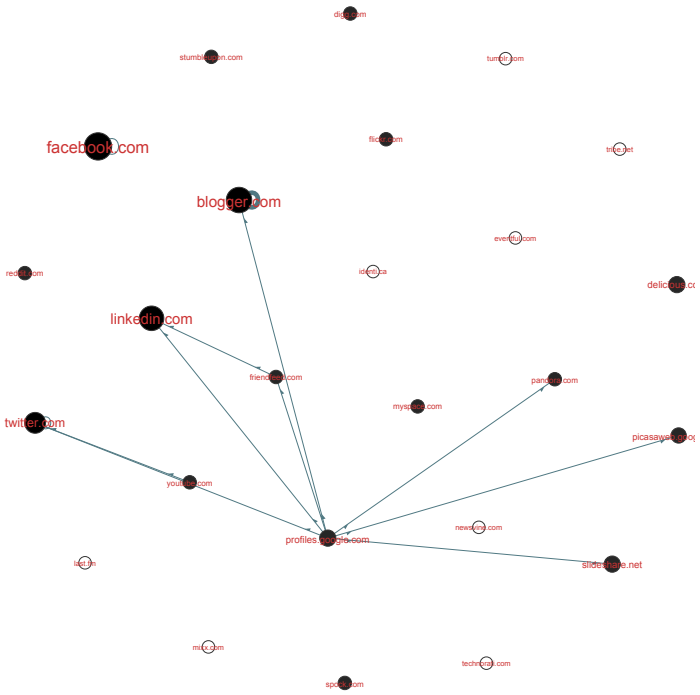

Figure 6: Truth Set, Services

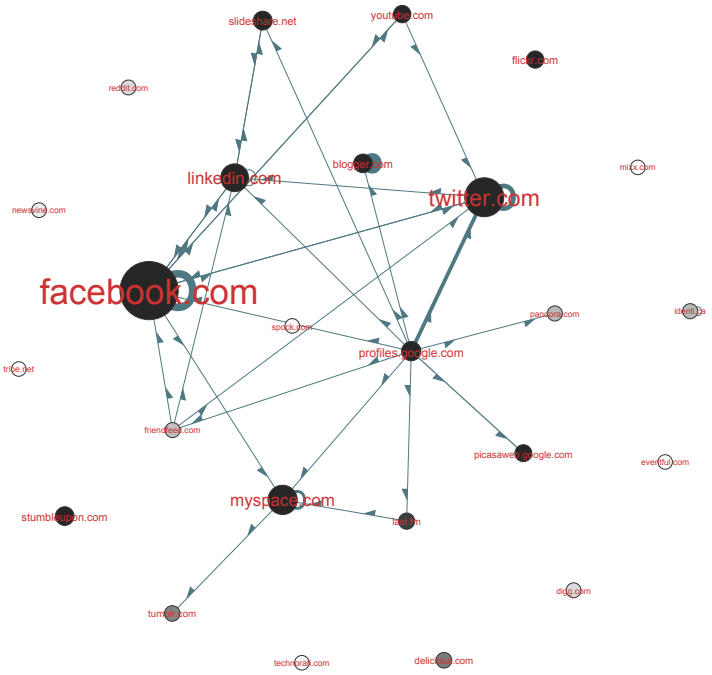

Figure 7: Data Set, Services

inspection is an average per site profile of about 10 friend links. Obtaining more friend links is desired as it could increase our precision, but this data is even harder to find, and sometimes not publicly available. This will be an area of future work to explore further.

The graph of Figure 8 contains 1866 weakly connected components and 1981 strongly connected components. The average clustering coefficient is 0.007. The average degree is 0.297 with an average In Degree of 0.149 and an average Out Degree of 0.149. The average path length is 2.052 and the network diameter is 6.

Figure 9, is the same graph as Figure 8, but without the links between users of Blogger. Notice in Figure 8 that the WS-DL research group is very highly connected while no other component in the graph shares that high rate of connectivity. This is due to the members of our research group being part of Blogger research blog which results in eight links to co-authors of the blog on our respective profile pages. By removing these links, we get a more accurate representation of just how our internal research group fits into the rest of the student body. Notice that while there is some connectedness, it resembles the connectedness and fits into other connected components of the graph seamlessly.

6. CONCLUSIONS

In this paper we presented a simple, accurate (precision of 0.863 and an F-measure of 0.654) counting method to disambiguate social media profiles for members of an organization. We do so with an unsupervised approach that uses keyword matching, community structure analysis, and extraction of web results and semantic data from the web. We tested our

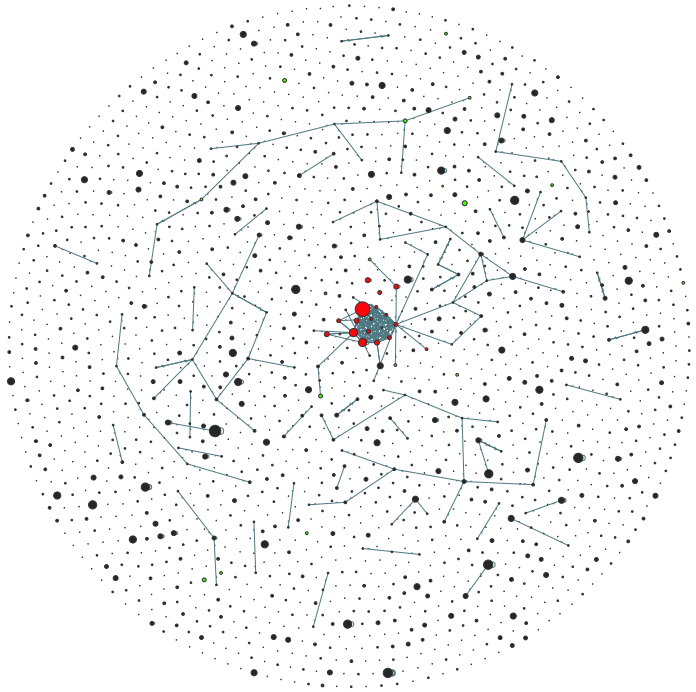

Figure 8: Data Set, User Accounts

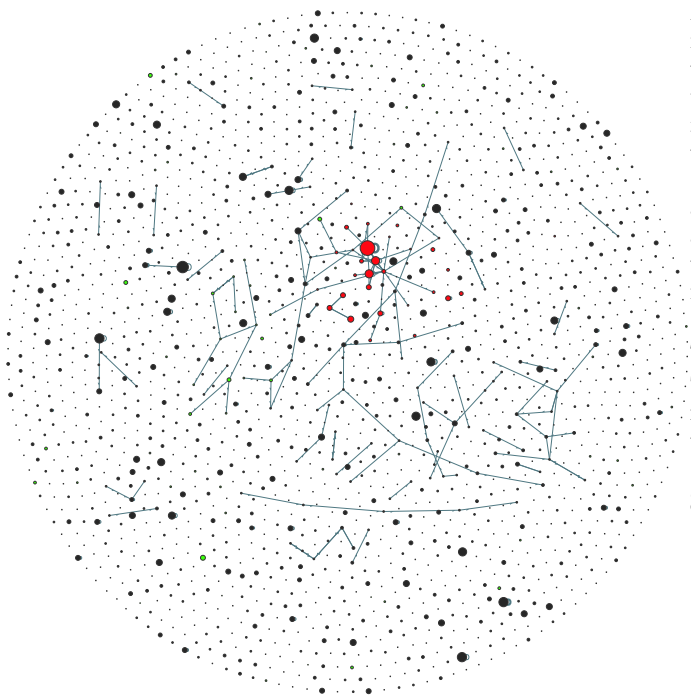

Figure 9: Data Set, User Accounts Without Blogger Links

approach with the known social media profiles of 22 members of our research group, then extrapolated the results to 2016 departmental Unix accounts.

The requirements stated in section 3 have been fulfilled as follows:

1. The approach is completely automated with the only human interaction being with the creation of the search query.
2. We have achieved a precision of 0.863.
3. We have achieved a recall of 0.526 and an F-measure of 0.654
4. Our approach is able to find profiles which are not indexed by current search engines by utilizing polling of social media sites for existence of profiles and through other means such as Google's Social Graph and Rapportive.
5. The approach uses non-traditional search mechanisms to achieve it's goals.
6. Only publicly available information was used; no privileged departmental information was used to track identities.
7. Our approach focuses on the profiles of 25 popular social media sites.
8. New social media sites can be added by changing the underlying Java code.

A number of improvements to this study could be made in the future, including verifying the results with a larger truth set, tuning the social media sites to remove unpopular sites and try to include other, emerging sites (e.g., `foursquare.com`), and expand our ability to map nicknames to their formal names for other than English language names (many of the students were foreign, especially in the graduate student body). A rather intriguing modification to the approach would be to use image processing facial recognition to cluster profiles into groups based on similarities of faces in profile pictures.

7. ACKNOWLEDGMENTS

Thanks to Dr. Matthew Rowe for providing background research into this process. This work supported in part by the NSF, Project 370161.

8. REFERENCES

[1] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins. Sic transit gloria telae: towards an understanding of the web's decay. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 328–337, New York, NY, USA, 2004. ACM Press.

[2] Z. Chen, D. Kalashnikov, and S. Mehrotra. Adaptive graphical approach to entity resolution. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 204–213. ACM, 2007.

[3] R. Nuray-Turan, Z. Chen, D. Kalashnikov, and S. Mehrotra. Exploiting web querying for web people search in weps2. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[4] B. On, E. Elmacioglu, and D. Lee. An effective approach to entity resolution problem using quasi-clique and its application to digital libraries. In *Digital Libraries, 2006. JCDL'06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on*, pages 51–52. IEEE, 2007.

[5] M. Rowe. *Disambiguating identity web references using social data*. PhD thesis, University of Sheffield, 2010.

[6] M. Rowe and F. Ciravegna. Disambiguating identity web references using Web 2.0 data and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2010.

[7] Y. Tan, M. Kan, and D. Lee. Search engine driven author disambiguation. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 314–315. ACM, 2006.

[8] P. University. Princeton university "about wordnet". `http://wordnet.princeton.edu`.

[9] C. van Rijsbergen. *Information Retrieval*. Someone Fake, Butterworths, London, 2nd edition, 1979.

[10] X. Wan, J. Gao, M. Li, and B. Ding. Person resolution in person search results: Webhawk. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 163–170. ACM, 2005.

APPENDIX

Table 4: Social Media Web Sites

Service	Description
blogger.com	A blog publishing service
delicious.com	A social bookmarking service
digg.com	A social news website
eventful.com	Search, track, and share information about events
facebook.com	A social network service and website
flickr.com	An image hosting and video service and online community
friendfeed.com	A social media service aggregator
profile.google.com	Aggregates social identities
identi.ca	Social networking and micro-blogging service
last.fm	A music recommendation service
linkedin.com	A business-oriented social networking site
mixx.com	Social networking, blogging, bookmarking and recommendations
myspace.com	A social networking website
newsvine.com	Collaborative journalism news website
reddit.com	A social news website
pandora.com	An automated music recommendation service
picasaweb.google.com	A photo sharing service
slideshare.net	An slide and file hosting service
spock.com	People and identity searching
stumbleupon.com	A discovery and recommendation service
technorati.com	Blog searching
tribe.net	Social networking sites
tumblr.com	A microblogging service
twitter.com	A social networking and microblogging service
youtube.com	A video sharing website