

Classification of Commercial and Personal Profiles on MySpace

William Gauvin^{*†}, Cindy Chen^{*}, Xinwen Fu^{*}, Benyuan Liu^{*}

^{*} Department of Computer Science
University of Massachusetts Lowell
Lowell, MA 01854, USA
{wgauvin, cchen, xinwenfu, bliu}@cs.uml.edu
[†] Symantec Research Labs

Abstract—Online social networks such as MySpace and Facebook have become popular platforms for people to make connections, share information, and interact with each other online. In an online social network, user publishing activities such as sending messages and posting photos, represent online interactions between friends. As more and more businesses use social networks as a means to propagate their “brand name” and distribute information about their product, a good understanding of user publishing characteristics is important for marketing analysis and aids in the ability to provide security measures for online social networks. In this work, we look at the implications of social networks with respect to commercial use. We are particularly interested in classifying commercial and personal profiles to protect the privacy and anonymity of individual users. We present an algorithm that uses online social network publishing relationships, such as publisher age distribution and usage patterns to construct a decision tree based classifier. The result is a C4.5 pruned decision tree which is applied to a Privacy-Preserving Data Publishing (PPDP) service to provide anonymity for online social network users.

I. INTRODUCTION

Online social networks have become an effective means for the sharing of commercial information on products, music, and entertainment. As such, the use of a social network profile has grown from a totally personal experience to sites that offer information for pecuniary gain. Many of these sites use “Branding”, which provides the means to associate specific products, such as NikeTM, with a specific symbol for a wide range of sports activity. Online social networks provide a medium to facilitate a large host of viewers with very limited company resources. In addition, OSNs attract a diverse community of users with a broad set of interests. Other commercial sites brand their music, TV show

or famous persona. For these commercial sites, friend publishing characteristics differ from those sites used for personal social networking.

The attraction of social networks has not gone unnoticed by marketers and sources with less than honorable intentions. As such, OSNs have also become a target for exploitation of naive members, their privacy and anonymity. The ability to identify commercial sites and provide a measure of anonymity is highly desirable to protect users against targeted advertisement, such as Facebook’s Beacon advertisement system. This research focuses on the protection of privacy and anonymity of individuals using online social networks, by first identifying commercial profiles and second, by defining a service that provides anonymity when personal profiles desire to correspond with commercial profiles.

A tool was developed to sample MySpace, uniformly at random, to collect user profile attributes (name, gender, age, location, friends, etc) and their corresponding blog information. Analysis shows that there are distinct profile and publishing characteristics, such as gender, age, and publishing attributes, which are unique to commercial profiles.

This paper will consider the specific attributes of commercial sites as compared to personal sites, and answer the following questions:

- *Membership Patterns*: What is the distribution and relationship between personal and commercial profiles with respect to the number of friends and the number of distinct publishers?
- *Age Patterns*: What is the relationship between age distribution with respect to personal and commercial profiles?
- *Gender Patterns*: Does gender play a role in the popularity of commercial profiles? If so, can this

relationship be characterized?

The answers to the questions above, results in a C4.5 decision tree whose accuracy varies from 92.25% to 96.4% depending on the attributes used. The ability and “cost” of obtaining these attributes will be examined. The algorithm is then applied to a Privacy-Preserving Data Publishing (PPDP) service to provide anonymity for online social network users. This service uses impersonation to allow an avatar to publish, on behalf of a personal profile, content to a commercial profile wall.

In the remainder of this paper we will present related work and background research on social networks in section II. In section III, we will describe the collection process, and our observations. In section IV, we describe the means used to derive the decision tree algorithm and the process used to validate the algorithm. In section V we outline an implementation, which provides privacy and anonymity for individual users. Finally, we provide our conclusions in Section VI.

II. RELATED WORK

Most of the related work in privacy protection is based on the study around Privacy-Preserving Data Publishing (PPDP) [9]. This approach seeks to hide user identity and prevent the leakage of sensitive data which could be used to impersonate an individual. The work uses anonymization to facilitate this approach. The publishing techniques proposed privacy-preserving tools such as customized web browsers and minimal information disclosure protocols for e-commerce activities, but does not specifically use the publishing characteristics of social networks to identify commercial profiles to anonymize the communication between the two.

There has been a flurry of studies on the measurement and analysis of various characteristics of online social networks. In [3] the authors looked at various statistics of male and female profiles in MySpace, including the friend degree distribution, user demographics, language and privacy preference. [2] study the relationship between friends and publishers and whether social links (social network friendships) are valid indicators of real user interaction. In our previous work [1], we observed that only a small fraction of the friends actually write to a user’s blog wall, we will explore this relationship relative to commercial profiles. [3] analyze word frequencies on MySpace wall posts, grouped by gender and age. Our previous work [1] analyzed publishing patterns pertaining to posted content such as hyper-links, images, and objects with respect to gender. We use these observations and, apply the techniques defined in [5]

for mail SPAM classification, to provide a means to identify commercial profiles of online services. We then introduce an anonymization technique to provide both privacy and anonymity for online social network users.

III. DATA ANALYSIS

In our previous work [1], we randomly sampled 6 million possible user profiles, the results of this process produced 614,970 public profiles of which 92,653 sets of blogs were obtained for a total number of 1,867,299 blogs. To determine the publisher characteristics, such as gender and age, 1,768,884 addition profiles were scanned. Using the public profile dataset, we visually inspected approximately 5,000 randomly selected profiles to classify these profiles as commercial or personal. This sub-set is used as the training data set for our analysis.

A. Gender Distributions

Female, male, and neutral profiles each constitute 52.8%, 37.7%, and 9.5% of the total valid IDs, respectively. Almost all (99.999%) gender neutral profiles are public as these are usually commercial profiles trying to attract people to visit their pages. Our results determined that 86% of the profiles that reported themselves as public, neutral profiles, where in fact commercial profiles, whereas only 7% of the male and female profiles were found to be commercial profiles. This is consistent with the observation that gender neutral profiles generally correspond to commercial profiles and is a leading indicator to ascertain whether a site is commercial or personal.

B. Member Profile Patterns

Figure 1 shows the age distribution of classified personal and commercial profiles. The majority of profile owners have ages within the range of 15 to 30. Beyond this range the distribution falls off rapidly, only to increase around age 69 and 100 due to a common practice among under age users to elude the age and privacy restrictions by reporting an exaggerated age[4]. An interesting observation is that most of the commercial profiles are located at age zero. Many commercial profiles do not provide the age nor gender attribute. This is a common trend which is exploited in the classification.

C. Publishers versus Friends

In an online social network, friend relationship does not necessarily reflect the real online interactions between users. Our datasets show an interesting relationship between the total number of friends of a user and the number of friends that actually publish on the user’s wall.

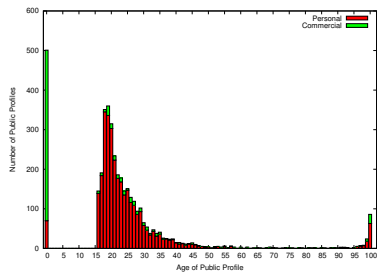


Fig. 1. Distribution of MySpace Public Accounts by Type

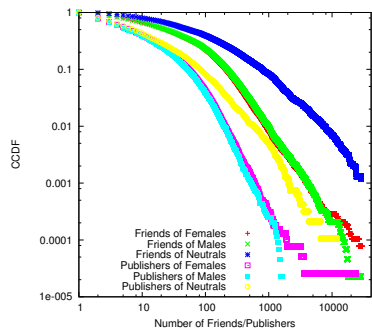


Fig. 2. Distribution of Number of Friends and Publishers

We observe that the distributions for male and female profiles exhibited in the complementary cumulative distribution function in Figure 2 have similar behaviors except at the tail, where the female distribution extends farther and the male distribution falls off sharply. The distribution for gender neutral profiles lies above the male and female distributions. This indicates that neutral profiles tend to have more friends and individual publishers than normal male and female profiles, but that their publishers do not publish as frequently. This trend is consistent with our observation that the neutral profiles are generally music band or other commercial profiles with a large number of friends that tend to publish once to comment on the music, movie, or product being promoted. We examined the number of distinct publishers versus the number of friends of MySpace users. As the number of friends increases, only a small fraction of the users actually interact with them through blogs. This is consistent with the observations made in [2] for Facebook and indicates that commercial, gender neutral profiles tend to have more friends than personal male and female profiles; however, an even smaller fraction of these friends actually write on their blog walls. A possible explanation for this difference is that many users befriend commercial profiles just to browse or receive information from these profiles. The relationship between friends and gender neutral commercial profiles

are less "active" than that between normal users.

D. Profile Owner and Publisher Age Difference

The age difference between publishers and profile owners differs significantly between commercial and personal profiles. Specifically, commercial profiles attract a more diverse audience, as such, the age difference is within 15 years of age, as compared with five years of age for personal profiles.

IV. CLASSIFIER

Three major considerations were taken into account during analysis; they are the ease in collecting the attributes, the actual classification technique used, and the ease of implementation. Several techniques are well known for data classification; they are the C4.5 decision tree, Bayesian classifiers, Neural Networks, and Support Vector Machines. In [5] a comparison is made of the techniques listed above for Email Spam analysis. The research concluded that the J48 decision tree proved to be one of the best for accuracy when using a limited set of attributes, as well as resulting in a binary tree, which is easy to implement. This insight proved true for the classification of commercial profiles as well.

A. Classification Criteria

The most important aspect of classification is the selection of the attributes used for analysis. Some of the attributes, *publisher age distribution* and *publisher versus friends ratio* require deep profile inspection, while others, such as *gender*, *account*, *age*, *friends* and *blogs* do not. The performance aspects are important when quickly determining whether a profile is used for personal or commercial use. The static attributes can quickly be ascertained from the destination profile with limited overhead. Deep inspection attributes require HTML scraping of the profile blogs and the owner's profile of those blogs. This information can be obtained in background processing, but not for real-time operations. Because of this factor, only static attributes were desired to be used in the implementation of the avatar. Using the attributes defined above, we identified 3 data-sets. The first data-set consists of the attribute set, *gender*, *age* and *friends*, the second attribute set consists of all static attributes, which included *gender*, *account*, *age*, *friends* and *blog counts*. Finally the third set contains all attributes, both static and deep inspection, to include the *publisher age distribution* and *publisher versus friends ratio*. Using WEKA, these data-sets were

used to compare the accuracy of the classifiers under consideration. As can be seen in Table I the J48 decision tree implementation resulted in the best accuracy given our dataset.

Classifier Comparison				
Classifier	NN	SVM	Naive Bayesian	J48
3 Attributes	87.14	91.29	88.7	92.25
5 Attributes	89.31	91.29	88.32	95.75
7 Attributes	89.40	91.34	86.22	96.42

TABLE I
CLASSIFIER COMPARISON

When comparing the result from the full attribute set and the static set, the full attribute set resulted in a solution with 66 leaves and a tree size of 127 before manual pruning, while the static set resulted in 57 leaves with a tree size of 109 before manual pruning. The precision lost was 0.67 percent, but the simplification in obtaining the data and implementation far outweigh this loss of precision.

B. J48 Classifier

Decision trees are commonly used to classify heterogeneous collections of data. They rely on the ability to order attributes into identifiable branches of a tree. J48 is an algorithm for generating C4.5 pruned or unpruned decision trees. This algorithm was designed by Ross Quinlan [7], [8] and is facilitated by the use of the WEKA tools [6]. Decision trees are created within the J48 algorithm by using information entropy on a set of training data. Data attributes are organized into subsets and the normalized information gain, measured by the difference in entropy, is used to measure these subsets to identify the optimum attributes used as nodes in the decision tree. When sub-setting, homogeneous splits can occur, producing an entropy value of 0. Data attributes are selected as the main attributes of the tree, resulting in a leaf node entry to derive a class specification. J48 is a recursive algorithm, which consolidates remaining data subsets, and continues the normalization information gain using the splitting process until all instances result in the same classification. Once created, the tree may be pruned to clarify the solution.

1) *Data Collection Sets*: Two data sets were derived from the original 614,970 public profiles collected by randomly creating a list of “analysis” and “holdout” sets. For these sets, each profile was manually inspected and classified to be either commercial or personal. During the manual inspection, it was determined that several

other categories are required to correctly describe the state of the profiles. For our study, any profile, which was once public, but changed to *private*, is considered for personal use. *Removed* profiles no longer exist, therefore could not be evaluated. *Reused* profiles were identified as profiles for which the name, gender, age, location, friends count and blog count were completely different from the automated scan. *Blocked* profiles are profiles that received a reset from the MySpace server during a connection attempt. Interestingly, these resets occurred for specific profiles, while others were easily viewed. These profiles were tested repeatedly at random intervals with the same results. Wireshark traces showed the server responded with a 302, “object moved”, and then issued a TCP disconnect. The last category was *undecided*. A conclusion for these profiles could not be ascertained for various reasons, for example, a page style rendered the content un-viewable. The result of the classification of 6,366 profiles resulted in 5,153 profiles available for analysis, with 788 commercial profiles and 4,365 personal. The analysis and “holdout” sets consisted of 3905 and 1248 profiles respectively. Of these, there were 667 commercial and 3238 personal profiles identified in the analysis set and 121 commercial and 1127 personal profiles identified in the hold out set. Table II outlines the remaining categories in the two sets.

TABLE II
ADDITIONAL CATEGORIES

DataSet	Private	Removed	Reused	Blocked	Undecided
Analysis	371	363	88	20	20
Holdout	181	145	10	2	10

2) *J48 Pruned Tree*: We have introduced a number of recognizable patterns with respect to gender, publisher distribution, age distribution and content types distribution for commercial and personal profiles. Some of these, such as gender, may be represented as a specific binary decision. Other, such as publisher distribution and age, have a higher range of values and need to be evaluated based on its relationship with other attributes associated with a profile. For the attributes identified as providing value when trying to classify a commercial or personal profile, the entropy $i(N)$, information impurity, for each was calculated using the formula below [7], [8], where $P(w_i)$ is the fraction of patterns at node N that are in the category w_i .

$$i(N) = - \sum_{i=1}^n P(w_i) \log_2 P(w_i)$$

This measurement is used to prioritize the decision tree and determine which nodes constitute the best analysis for each attribute for the levels of the decision tree. Again, for simplicity, a monothetic tree is desired. The class entropy was determined to be 0.691. Separating the remaining attributes into sub-sets, it was determined that age has the best entropy with a value of 0.583, while gender resulted in entropy of 0.578 and the number of friends following at 0.531. This analysis gives us a good indication on the priority of the attributes we are considering. Using the attributes with the highest information gain from the above calculation, *gender*, *age*, and *friends* were used to produce a decision tree of size of 9, with 5 leaves and correctly classified 92.25% of the instances. The stratified cross-validation is 91.8% and a loss of precision of 3.5% compared to the static attribute set listed above. This decision tree result is listed in Figure 3.

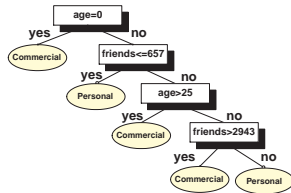


Fig. 3. J48 Decision Tree

Important to note is the absence of the gender attribute, which ranked high in our manual analysis, but is removed when separating the attributes in subsequent subsets by the J48 classifier. Next the ‘hold out’ set was used to validate the algorithm. Using WEKA, 94.39% were correctly classified. This “over-simplified” version of the algorithm was used in the implementation of the avatar classifier.

V. PRACTICAL APPLICATION

To circumvent the loss of privacy and maintain anonymity, the classifier algorithm in section IV is used to ascertain when commercial profiles are the target of a user publishing activity; and redirects the publication to an anonymous privacy avatar, which post the desired content on behalf of the personal profile. A privacy avatar is a valid on-line social network profile, which is created with the specific intention of hiding a personal profile user’s identify. The avatar profile is anonymous, to include no reference to a specific name, age or location. In addition, the profile image associated with the profile is opaque. The avatar is used as an intermediate focal point when publishing to commercial sites.

Cross-references are maintained in the avatar database to associate personal profile content with an avatar post. This association is used to “link” the two profiles and allow valid content to be delivered to the personal profile from the commercial site, if desired. The implementation of the avatar provides the additional benefit of being a funnel for SPAM, Data Loss and Antivirus protection and can act as a sensor for new JavaScript malware analysis [10].

A. Overview

The privacy avatar is implemented as a C#, .NET multi-threaded transparent proxy, which provides the ability to intercept, inspect and impersonate social network communication. This overview will focus on the details and implementation of a client only solution, but this solution can easily be extended to a cloud-based strategy. Figure 4 generalizes the overall architecture of the avatar privacy system.

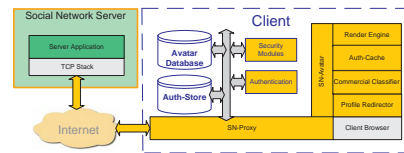


Fig. 4. Privacy Avatar Architecture

The main components of the system are the online social network server, avatar service and client browser. The online social network service is unmodified and performs the actions normally associated with these services. The avatar solution includes a transparent proxy, classifier and avatar engine. The transparent proxy is used to provide an injection point into the network stream when a personal profile is interacting with an Online Social Network (OSN) service. It intercepts, evaluates and redirects personal profile interaction directed to commercial profiles. The evaluation process uses a classifier to ascertain whether a destination profile is a commercial profile. Such posts are redirected to the avatar engine for all interaction with commercial profiles. Besides the *transparent proxy*, *avatar engine*, and *commercial classifier*, the other specific components of the avatar solution are the *Rendering Engine*, *Authentication* (used to automatically log-in to the avatar account), and various security filters such as anti-spam, antivirus, malware detection and data loss prevention.

B. Avatar Masquerading and Impersonation

The avatar proxy provides the ability to intercept client publishing request to commercial profiles and re-

publish the submission using an anonymous avatar to provide anonymity for the personal profile. The classifier ascertains whether a destination profile is either commercial or personal for a specific publishing activity. A database of previous classifier results are recorded and cached. Responses to the posting from the commercial profile are directed to the avatar, which in turn, scans the response for security violations, filters based on policy, and re-distributes the posting to the personal profile if all criteria are met. Commercial profile content is stored in the avatar database, in order to render the content when the personal profile is viewing their wall. This content appears as though it was directly published by the commercial profile, when in fact it is being injected by the avatar. The rendering of content on a commercial profile wall by a personal profile, using an avatar, is recorded in the avatar database, as well. The database entities are used to ascertain the blogs for which a personal profile is specifically responsible for, and allows the avatar to fix up the rendering of content for which the personal profile is the originator of, when viewing the commercial profile wall.

C. Avatar Publishing Operations

The primary component of the avatar is a C# Web-Browser control object, which provides the means to automatically, and dynamically emulate avatar publishing request on behalf of a personal profile. When a post, destined to a commercial profile, is identified, the transparent proxy redirects the request to the avatar in the form of a request object. This object includes the destination profile ID and content to be published. The avatar dynamically navigates to the commercial profile using the “navigate” method of the WebBrowser object. The avatar then uses the HTMLDocument DOM object, to inject the required publishing content and invokes the post action to initiate the publishing of this content using the avatar, on behalf of the personal profile. When the personal profile is viewing their wall, the rendering engine within the avatar is used to inject commercial content that has passed filtering. MySpace uses JASON objects as its method to forward request and receive responses. These responses are wrapped in an HTTP response in compressed form. The rendering engine decompresses the response and injects the additional content, compresses the response and updates the HTTP content header size. It then forwards the response to the client browser for rendering. The results are complete commercial post entries, which appear as though they were rendered by the OSN server.

VI. SUMMARY

The classification of personal and commercial profiles is important when applied to Privacy-Preserving Data Publishing (PPDP) services, used to provide anonymity for online social network users. The classifier is a decision tree used to identify a profile as being either commercial or personal. The decision tree uses profile attributes which exhibit distinct patterns between personal and commercial profiles, to include age, gender and publishing relationships. The result of the classifier yields a binary tree with a degree of accuracy of 92.25% to 96.42, depending on the attributes selected for use. To circumvent the loss of privacy, an avatar solution is presented. Using the results of the classifier attributes, an avatar transparent proxy can detect and post, on behalf of a personal profile, endorsements, without losing privacy and anonymity. In addition, commercial content may be securely filtered and rendered to personal profile walls. This work has provided a number of key insights which can be used to provide solutions for existing problems or extended in future work.

REFERENCES

- [1] W. Gauvin, B. Ribeiro, B. Liu, D. Towsley, and J. Wang, “Measurement and gender-specific Analysis of User Publishing Characteristics on Myspace,” in *IEEE Network*, September/October 2010.
- [2] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, “User interactions in social networks and their implications,” in *ACM EuroSys 2009*, 2009.
- [3] J. Caverlee and S. Webb, “A large-scale study of MySpace: Observations and implications for online social networks,” in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2008)*, Seattle, WA, April 2008.
- [4] H. Nigam, “Joint Statement on Key Principles of Social Networking Sites Safety,” *blogs.myspace.com/hemanshunigam*, APPENDIX A. 1, April 2008.
- [5] S. Youn, D. McLeod “A Comparative Study for Email Classification,” *Advances and Innovations in Systems, Computing Science and Software Engineering*, pp.387–391, 2007.
- [6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten “The WEKA Data Mining Software: An Update,” *SIGKDD Explorations*, Volume 11, Issue 1., 2009.
- [7] R.J. Quinlan “Learning with Continuous Classes,” *5th Australian Joint Conference on Artificial Intelligence, Singapore*, pp.343–348, 1992.
- [8] R.J. Quinlan “C4.5 Programs for Machine Learning,” *Morgan Kaufmann Publishers*, 1993.
- [9] B. Fung, K. Wang, R. Chen, P. Yun “Privacy-Preserving Data Publishing: A Survey of Recent Developments,” *ACM Computing Surveys (CSUR)*, Volume 42, Issue 4, June 2010
- [10] M. Cova, C. Kruegel, G. Vigna “Detection and Analysis of Drive-by-Download Attacks and Malicious JavaScript Code,” *Proceedings of the International World Wide Web Conference (WWW)*, Raleigh NC, April 2010