Are Your Friends Who They Say They Are?: Data Mining Online Identities

By Roya Feizy, Ian Wakeman, and Dan Chalmers

How sure are you that your friends are who they say they are?

In real life, unless you are the target of some form of espionage, you can usually be fairly certain that you know whom your friends are because you have a history of shared interests and experiences. Likewise, most people can tell, just by using common sense, if someone is trying to sell them on a product, idea, or candidate. When we interact with people face-to-face, we reevaluate continuously whether something just seems off based on body language and other social and cultural cues.

These identity validation questions have a long history in computer science and translate directly to the pervasive computing context, where there is a widespread view that access control mechanisms will use some form of computational trust [8]. One example of this paradigm is the set of social networks embodied in Web sites such as MySpace and Facebook. If a person can show proof that he or she is responsible for an online identity through standard public key cryptography, then his or her information and relationships can be used to calculate a level of trust.

The millions of social network users and billions of connections between them make it non-trivial to formalize an automated approach to differentiate fact from fiction in self-described identities online. An identity may be part of a role-playing game [1] or it may be an impersonation, either for play or more nefarious purposes, such as fraud. However, each of these identities still has associated profile data and is embedded within a social network.

How can we be sure with whom we are interacting and whether these individuals and groups are being truthful in the online identities they present to the rest of the community? What tools and techniques can be used to gather, organize, and explore the available data for informing the level of trust that should be granted an individual? Can we verify the validity of the identity automatically, based on the displayed information?

To tackle these questions, we use a machine learning approach to look at traces of people's identities left behind on online social networking sites to evaluate the validity of those identities. We train classifier-based models on profiles with known identities (real or fake). We also use data mining techniques and social network analysis to extract significant patterns in the data and network structure and improve the classifier during the cycle of development.

We evaluate our algorithm on 2.2 million user profiles' features collected from MySpace. Our results indicate that by utilizing people's online, self-reported information, network of friends, and interactions, we are able to provide evidence for deciding the level of trust with which to imbue individuals in making access control decisions in a manner that is both accurate and scalable.

Social Network Data Collection

To obtain our sample data set, we customized a robust crawler to accumulate personal and relational information from MySpace profiles within three main categories: 1) public (personal pages), 2) private (pages with limited biographical data) and 3) bands (data related to musical artists). (See Table 1.)

	Unknown (valuation set)		Known (real/fake) (training & test set)		Total
	seed	friends	seed	friends	
Public	113,969	1,101,032	701	9,389	1,225,091
Band	19,908	181,371	219	2,147	203,645
Private	68,958	725,995	73	3,430	798,456
Total	202,835	2,008,398	993	14,966	2,227,192

Table 1: Number of collected profiles by each category.

Each seed identity was chosen by selection of a random FriendID (MySpace members' unique number). We then crawled pages up to a depth of two degrees (link to the friends and the friends of friends), targeting the top 40 friends of each individual.

We applied a qualitative study to manually identify the true identity of three types of users for our classifier training data:

- Real (popular): official profiles representing famous people. These are obviously well-connected profiles, which might affect the experimental results; therefore we collected a number of local users. (417 participants).
- Real (local): current students at University of Sussex who responded to a survey (118 responses from 2,019 emails) and verified that their profiles belong to them, and rated their level of honesty.
- Fake (impersonator): users who fabricated real persona with almost the same information such as name and pictures. We determined fakes manually, for instance by knowing of another real profile for the same person (457 participants).

The known data (real or fake) was used as the training and testing set, while the remaining unknown dataset was used to investigate appropriate pre-processing algorithms for the classifier (Table 1).

Profile Personalities

To aid the machine learning classifier, we developed a series of attributes to describe each individual profile. A preprocessing algorithm to derive a set of personality features from the raw profile data was devised. The set of personality features includes:

- expressive/anonymous,
- valid/fantasy,
- active/inactive,
- positive/offensive,
- popular/isolated,
- sociable/unsociable, and
- traceable/untraceable.

Our approach to labeling the training data was empirically driven, experimenting with formulae to best match the collected data. The features are determined using a mixture of ad-hoc automated techniques, ranging from checking the validity of the address to comparing the terms and language used against a list of known terms.

For each feature pair, a profile is awarded a normalized score between 0 and 1. Using mixtures of these features, we were then able to classify each profile along three scales. (See Figure 1.)

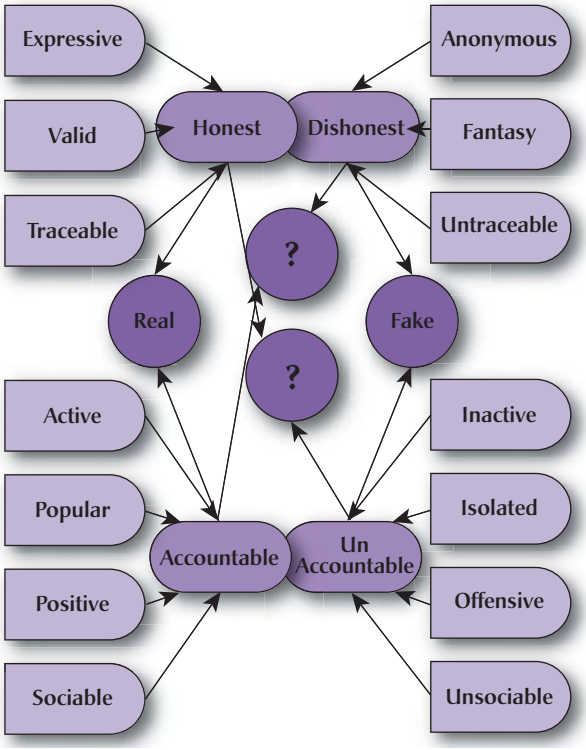

Figure 1: Identity model based on personality factors.

1. *Honest/dishonest.* It can be argued that there is a trade-off between privacy and honesty in online presentation dependent upon context [10, 16]. We define an honest person as one whose information is valid (exists and is reasonably acceptable); traceable (includes a Web address, school name, and photograph), and expressive (a count of information revealed).
2. *Accountable/unaccountable.* Accountable/unaccountable can be defined based on a user's level of activity (blog and group membership), sociability (number of comments), popularity (number of friends and view number), and positive use of language (which we later see is not a strong indicator of validity).
3. *Real/fake.* We define a fake profile as one intended to make people believe that the profile belongs to some real person who has no actual connection to the profile. On the other hand, a real profile is one controlled by the identity presented.

Social Network Analysis

Because social networks are examples of small world networks [13], the community can be modeled as a network $N=I, F$, where I represents an individual or node, and F represents a friend's link or edge.

Social network analysis can be used to describe the properties of this network structure as well as characteristics about a specific individual in that network. These include a profile's connectivity and the amounts and types of interactions with other members of the community which can reveal information about the validity of an identity.

To capture this information in a form that can be used by our classifier, we analyzed measurable characteristics including the out-degree, in-degree, overlapping (mutual friends), centrality, and isolation of nodes which were tagged as accountability attributes, to identify a relationship within these properties and the type of identity. We also measured the similarity criteria for both self-described data and extracted personality factors between individuals and their network of friends.

These data were generated by incorporating the identity features of the top 40 friends within the system. Our data set initially contained more than 4.8 million profiles, which reduced to 2.2 million nodes with 2.4 billion edges between them after removing mutual friends. This suggests the probability that a friend of a friend will become a friend is much higher than a stranger becoming a new friend.

Our analysis revealed that the sample network employs many high-degree connections, which strongly clustered with an average of 1,010 friends for public profiles and 5,792 for band profiles. Real-popular nodes showed a high out-degree distribution, although the out-degree analysis alone was unable to verify a fake person from real-local. From this analysis, we defined the popularity factor in our classifier as the centrality measurement, or the accumulation from out-degree and in-degree distribution.

The distribution of training node (known profiles) positions within the network structure shown in Figure 2 illustrates several key observations. The real-popular nodes are more closely linked to each other,

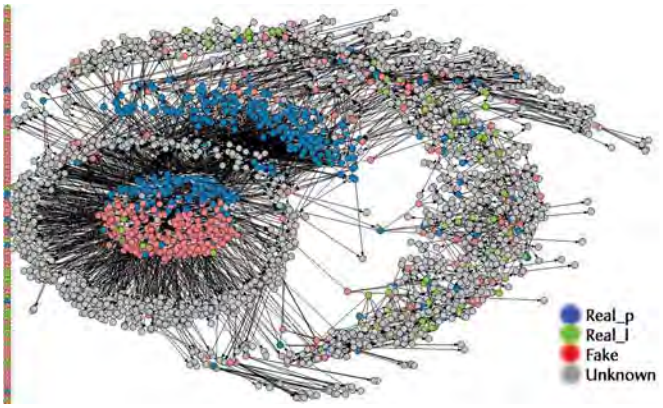

Figure 2: The network structure of training set.

Friendship also relies on some degree of similarity, meaning the characteristics of individuals should be, on some level, similar to those people who connect with them. This similarity measurement reveals information about the context of links between users and the correlation between their identity types.

The following formula was used to compare I, which represents the value of personality features for individuals, with F, which indicates the average friend's value of personality attributes. The average similarity was calculated by dividing the minimum value by the maximum value between individuals and their friends. (See Equation 1.)

$$Similarity\ (I,F) = \frac{min(I,F) * 100}{max(I,F)}$$

(1)

$$I = \sum_{a=1}^{n} i(a) \qquad F = \frac{\sum_{i=1}^{f}(\sum_{a=1}^{n} f(a))}{f}$$

This formula allowed us to uncover which identity elements are more important in choosing a friend. Our analysis of the pre-classified attributes showed that the traceability, validity, and being positive are not as important as being active, sociable, and popular. This similarity measurement reveals that people confer more value in accountability than honesty due to less honesty being a less visible characteristic.

Reality Algorithm

Using an individual's personality attributes and their relation within the larger social network, we were able to compute our reality algorithm for determining the likelihood that an identity is valid. In our identity model, $R(x)$ refers to a real person, $F(x)$ to a fake persona, and a represents each attribute extracted from their profiles. Our algorithm calculated values for each, indicating the likelihood that an identity is either real or fake based on the level of accountability and honesty.

To calculate $R(x)$, the summation of honesty $H(a)$ and accountability $A(a)$ values are added to the squared average of top friends' attributes. $F(x)$ is calculated based on the dishonesty $D(a)$ and the unaccountability $U(a)$ values. Top friends' personality values are counted in the weighting schema as from our experiment on the similarity attribute.

$$R(x) = \sum_{a=0}^{n} H(a) + A(a) + \sqrt{\frac{\sum_{i=1}^{f}\left(\sum_{a=0}^{n}(H(a) + A(a))\right)}{f}}$$

$$H(a) = avg(expressive + valid + traceable)$$

$$D(a) = avg(anonymous + fantasy + untraceable)$$

(2)

$$F(x) = \sum_{a=0}^{n} D(a) + U(a) + \sqrt{\frac{\sum_{i=1}^{f}\left(\sum_{a=0}^{n}(D(a) + U(a))\right)}{f}}$$

$$A(a) = avg(active + popular + sociable + positive)$$

$$U(a) = avg(inactive + isolated + unsociable + offensive)$$

We learned that the choice of friends has an influence on the individuals' determined type of identity. We took the square root of friends' attributes because they are less significant compared to the individual's attributes. (See Equation 2.)

Classifier Experiments

To identify patterns within our data and to improve our classifier model in parallel, we used a supervised learning approach to train and test our classifier model. We evaluated four classifier models to classify data more efficiently and determine which would operate most effectively given our problem definition: decision tree, rule learner, Naive Bayes, and nearest neighbor [21].

Given the large size of our sample population and the number of features used to describe each of the data points, we first used principal component analysis [22] to reduce the amount of dimensions required to cluster the data prior to attempting learning. This redundancy technique examines the correlation between features within the training data set and generates the main components with minimal loss of information. The result of this analysis indicates which factors or components are most significant when examining personal information to predict the truth about identities.

Using the attributes that were identified as the principle components, two-thirds of our known data, consisting of the both raw data and extracted personality attributes, was used as the training set (to build a model), and one-third is used as the test set (to measure the model). Several validation schemes exist that can be used to estimate the performance of a learner, such as simple validation, regression performance, and T-test.

> **"Despite the important role of honesty in social interaction, the honesty value (quality in content) has less effect than accountability value (quantity in interaction) for identity prediction in MySpace."**

In our experiments, we applied the cross-validation operator [21], which evaluates the learning method from the training set and applies the average absolute and squared errors to the test set to predict the unknown labels.

Determining performance accuracy of the learner produces a confusion matrix [21], which is an evaluation technique to factor a matrix of true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN) values, where:

- TP is correct classification of correct data: real correctly tagged as real
- TN is correct classification of incorrect data: fake correctly tagged as fake
- FP is incorrect classification of incorrect data: fake incorrectly tagged as real
- FN is incorrect classification of correct data: real incorrectly tagged as fake.

The metrics for performance evaluation can be calculated as:

$$Accuracy = TP+TN/TP+FP+FN+TN,$$

where

$$Precision = TP/TP+FN \text{ and } Recall = TP/TP+FP.$$

Table 2 demonstrates the performance of pre-classified data using the nearest neighbor method. The highlighted cells represent the TP and TN prediction with the average accuracy performance at 84.6 percent.

Accuracy: 84.59%		Actual Identity			
Predicted	real-popular	real-local	fake		class precision
	real-popular	377	19	26	89.34%
	real-local	15	62	31	57.41%
	fake	25	37	400	86.58%
	class recall	90.41%	52.54%	87.53%	

Table 2: Confusion matrix of Nearest Neighbor learner over training dataset.

We validated different learners for both original data (profile's content such as age, gender, location) and pre-classified data (extracted personality factors such as valid, popular, traceable) in order to achieve more accurate precision. Evaluating both of these sets allows us to compare the average performance improvement across all three inputs for the machine learning models.

Our results showed that the overall performance over pre-classified data is higher than using the original data, while incorporating social network data improves performance yet further. (See Table 3). Although the diversity of information in pre-classified data is less, it's much faster and the prediction performance is more effective than using the original data by 83.65 to 65.98 percent.

	Decision Tree	Rule Learner	Nearest Neighbor	Naïve Bayes	Overall Accuracy
Original data	69.25%	66.63%	67.05%	60.99%	65.98%
Pre-classified data	86.10%	85.89%	84.59%	78.03%	83.65%
Learner accuracy	77.68%	76.26%	75.82%	69.51%	

Table 3: Average learner performance comparison when using original data and pre-classified data.

Real vs. Fake

Our results reveal three important implications. First, they allow us to clarify our assumption that the levels of honesty and accountability have a strong correlation when determining real versus fake personas. As shown in Figure 3, the real nodes are more often associated with both higher accountability and honesty, while fake users have lower values for both attributes.

In our identity model (Figure 1), we identified the four possible types of identity representation: honest and accountable (HA), dishonest and unaccountable (DU), honest and unaccountable (HU), dishonest and accountable (DA). The fraction and frequency of each type of identity representation are shown in Figure 4.

The results prove our assumptions that:

1. HA highly correlated with real-popular users;
2. DU highly correlated with fake users;
3. HU mainly correlated with real-local users; and
4. DA mainly correlated with real-local and fake users.

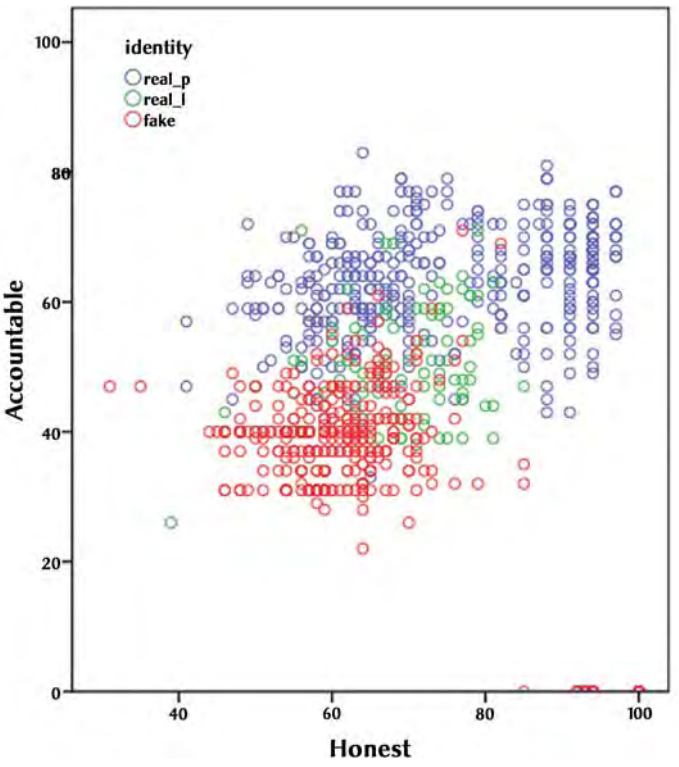

Figure 3: Relationship between honesty and accountability to determine the type of identity.

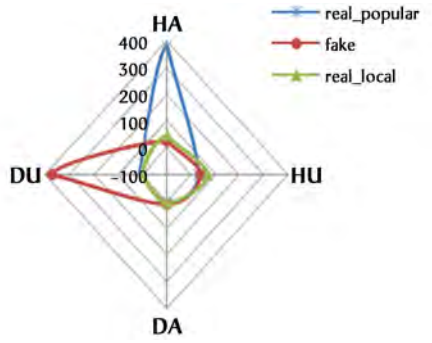

Figure 4: Four-dimensional identity representation.

Finally, our results show that, despite the important role of honesty in social interaction, the honesty value (quality in content) has less effect than accountability value (quantity in interaction) for identity prediction in MySpace.

Friend or Foe?

Our work draws upon research areas in computer sciences, statistics, sociology, and psychology. Previous work similar to ours has looked at the validity in self described online identities [6, 18]. Some of these studies showed promise in predicting personality traits with high accuracy [7], but were not able to clarify if the predicted traits are real or fabricated.

The social network analysis we performed was inspired by previous studies on understanding properties of social network structures [3], especially connectivity [11], interactions [4, 14, 15], and behavior

[9, 19]. These works focused on elements of sociability, demographics, network characterizations, privacy implications, and similarity measures within social networking. Our work takes the next step by attempting to determine the validity of an identity on an online social community by mining self-described data.

The data mining techniques in our approach have a long history of use in crime detection [17], intrusion detection models [5], and lie detection [20, 12]. Our algorithm adds to the already existing set of tools that can be used to identify forms of deception in nonverbal communication [2].

Although our approach is feasible and our results are promising, there remain many topics to investigate in future work. First, having more data would result in greater accuracy. Second, a comparison of the effectiveness of the present approach against other social networking sites such as Facebook could shed light on its generalizability and robustness. Third, if a user is really attempting to pass off a fake identity, would the classifier be effective in detecting it? Finally, it would be interesting to experiment with the resiliency of our classifier to carefully crafted identities and investigate how this classifier could be used as evidence in computational trust systems.

Biography

Roya Feizy is a PhD student at the University of Sussex. She holds degrees in Applied Mathematics from Azad University in Iran and Multimedia and Computer Science achieved from Middlesex University. Her research interests include identity and online social networking, specifically looking at how individuals present themselves online, whether they are real or fake, and the type and amount of information they are willing to disclose.

References

1. Boyd, D. 2007. Why youth (heart) social network sites: The role of networked publics in teenage social life. In *Youth, Identity, and Digital Media*, Buckingham, D., Ed. MacArthur Foundation Series on Digital Learning. MIT Press, Cambridge, MA.
2. Burgoon, J., Adkins, M., et al. 2005. An approach for intent identification by building on deception detection. In *Proceedings of the Hawaii International Conference on Systems Science* (*HICSS'05*).
3. Casciaro, T. 1998. Seeing things clearly: Social structure, personality, and accuracy in social network perception. *Social Netw. 20*. 331-351.
4. Caverlee, J. and Webb, S. 2008. 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*).
5. Dokas, P., Ertoz, L., et al. 2002. Data mining for network intrusion detection. In *Proceedings of the NSF Workshop on Next Generation Data Mining*.
6. Donath, J. and Boyd, D. 2004. Public displays of connection. *BT Technol. J. 22*, 4.
7. Hu, J., Zeng, H., Lin, C., and Chen, Z. 2007. Demographic prediction based on user's browsing behaviour. In *Proceedings of the 16th International Conference on World Wide Web*.
8. Kagal, L., Finin, T., and Joshi, A. 2001. Trust-based security in pervasive computing environments. In *IEEE Comm*.
9. Maia, M., Almeida, V., and Almeida, J. 2008. Identifying user behaviour in online social networks. *In Proceedings of the 1st Workshop on Social Network Systems*, ACM.
10. Mazar, N., Amir, O., and Ariely, D. (2007). The dishonesty of honest people: A theory of self-concept maintenance. *J. Market. Resear. XLV*. 633-644.
11. Mislove, A., Marcon, M., et al. 2007. Measurement and analysis of online social networks. In *Proceedings of the 5th ACM/USENIX Internet Measurement Conference* (*IMC'07*).
12. Mundinger, J. and Le Boudec, J. 2005. The impact of liars on reputation in social networks. In *Proceedings of Social Network Analysis: Advances and Empirical Applications Forum*.
13. Newman, M., Watts, D., and Strogatz, S. 2002. Random graph models of social networks. *Proc. Nat. Acad. Science*. 2566-2572.
14. Ryberg, T. and Larsen, M. C. 2008. Networked identities: Understanding relationships between strong and weak ties in networked environments. *J. Comput. Assist. Learn. 24*. 103-105.
15. Shrivastava, N., Majumder, A., and Rastogi, R. 2008. Mining (social) network graphs to detect random link attacks. In *Proceeding of the 24th International Conference on Data Engineering* (*ICDE'08*).
16. Somanathan, E. and Rubin, R. 2004. The evolution of honesty. *J. Econom. Behav. Organiz. 54*. 1-17.
17. Thongtae, P. and Srisuk, S. 2008. An analysis of data mining applications in crime domain. *Computer and Information Technology Workshops*.
18. Toma, C. L., Hancock, J. T., and Ellision, N. B. 2008. Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality Social Psych. Bull. 34*, 8. 1023-1036.
19. Tufekci, Z. 2008. Can you see me now? Audience and disclosure regulations in online social network sites. *Bull. Science Technol. Society 28*, 1. 20-36.
20. Whitty, M. T. 2002. Liar, liar! An examination of how open, supportive and honest people are in chat rooms. *Comput. Human Behav. 18*. 343-352.
21. Witten, I. H. and Frank, E. 2000. Data Mining: Practical machine learning. In *Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA.
22. Zou, H., Hastie, and T., Tibshirani, R. 2006. Sparse principal component analysis. *J. Computation. Graph. Statistics 15*, 2. 265-286