Distingushing Fact and Fiction: Data Mining online Identities

Roya Feizy[1]

Informatics Department
University of Sussex
Brighton, UK

Ian Wakeman , Dan Chalmers[2,3]

Informatics Department
University of Sussex
Brighton, UK

Abstract

As social technologies mature more people leave traces of their identity on online social networking sites. Such traces can be used as evidence in deciding the level of trust with which to imbue individuals in making access control decisions. However the online profile may not represent their true self. It is non-trivial to formalize an approach to differentiate the fact from fiction in self-described identities online. This study aims to investigate how people present themselves through their online social space by analyzing their self-reported information and through their network of friends and interactions. We propose a novel classifier model and evaluate its performance to predict the personality and validity of user's identity. We tested our algorithm on a large set of profiles' features collected from the popular social site MySpace. Using several data mining algorithms we trained a set of profiles with known identities (real or fake) to evaluate the accuracy of the classifier. Other approaches such as principal component analysis (PCA) and social network analysis (SNA) have been employed to extract a significant pattern in data and network structure and improve the classifier in the cycle of development.

Keywords: Identity, Online Representation, Social Network, Data Mining.

1 Introduction

There is a widespread view that the access control mechanisms for pervasive computing systems will use some form of computational trust [10]. One source of evidence for use in computational trust decisions can be the set of social networks embodied in web sites such as MySpace and Facebook. If a person can show proof that they are responsible for the creation of the online identity, e.g. through the use of standard public key cryptography, then the information and set of relationships on display

[1] Email: r.feizy@sussex.ac.uk
[2] Email: ianw@sussex.ac.uk
[3] Email: d.chalmers@sussex.ac.uk

can be used in calculating the level of trust. The question then arises as to whether we can verify the validity of the identity automatically, based on the displayed information. In this paper, we investigate whether machine learning techniques can be used to develop a classifier for the validity of social network identities.

Current social networking web sites are used for a variety of purposes, such as maintaining contact with friends, promoting musicians and other performers, accumulating prestige and so forth. However, the identities are not always a truthful representation of the real person - people may hide their real location or age for privacy reasons - the identity may be part of a role playing game [2]. Indeed, the identity may be an impersonator, either for playful mischief, or for more nefarious purposes, such as fraud. Each of these identities has associated profile data, such as personal, professional and relational, and is more or less embedded within a social network. Can this data and position within a social network be used to determine the validity of the identity automatically?

We collected 2.2 million profiles by spidering MySpace and determined the validity of a subset of about 1000 identities using qualitative techniques such as inspection and an email survey. We then experimented with a number of ways that the data can be used to validate an identity. We present interesting aspects of our results, and the resulting preprocessing strategy and classification engine below, with experimental results that demonstrate its quality.

2 Modeling Classifier

2.1 Data Collection

We have customized a robust crawler to accumulate personal and relational information of MySpace profiles within 3 main categories of public (personal pages), private (pages with limited biographical data) and bands (with music band related data) (see Table.1). Each seed identity is chosen by selection of a random 'FriendID' (MySpace members' unique number). We then crawled pages up to a depth of 2 degrees (link to the friends and the friends of friends), targeting the top 40 friends of each individual. To provide a training set for the classifier, we applied a qualitative study to manually identify the true identity of three types of users:

- `Real (popular):` official profiles representing celebrities (with 417 participants). These are obviously well connected profiles, which might affect the experimental results; therefore we collect a number of local users.
- `Real (local):` current students at University of Sussex who responded to our survey questionnaire (118 responses from 2019 emails) and verified that the profile belongs to them and rated their level of honesty.
- `Fake (impersonator):` users who fabricated 'real' profiles 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, whilst the remaining unknown dataset was used to investigate appropriate preprocessing algorithms for the classifier, and in other work reported in subsequent papers.

Table 1. Number of collected profiles by each categories

	Unknown (Evaluation set)		**Known (Training, Test set)**		
	Seed	Friends	Seed	Friends	**Total**
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

2.2 Personality Factors

As will be shown in Section 3, the raw identity traits (such as age, gender, location and so on) do not perform well in machine learning evaluation. We have therefore used a classification of the traits to preprocess the raw data into features. Our aim is to classify each profile along the following scales:

`Honest/ dishonest:` it can be argued that there is a trade-off between privacy and honesty in online presentation dependent upon context [12,18]. We define an honest persona when the information is 'valid' (exists and reasonably acceptable); 'traceable' (includes traces such as web address, school name and picture); and 'expressive' (a count of information revealed).

`Accountable/ unaccountable:` this can be defined based on the level of 'activity' (such as 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).

`Real/ fake:` we define fake as providing profile information intended to mislead people into believing that this is a profile of some real person, who has no connection to the profile, whilst real is a profile controlled by the identity presented.

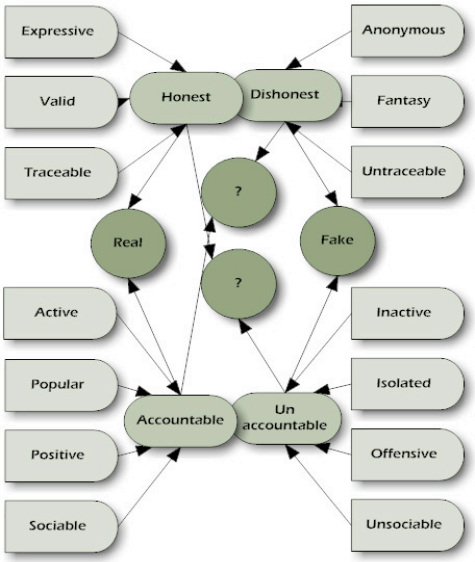


Fig.1. Classification model based on profile features

We have therefore devised algorithms to process the raw profile data into the following features: expressive/anonymous; valid/fantasy; active/inactive; positive/offensive; popular/isolated; sociable/unsociable; and traceable/untraceable (Figure.1). The approach has been 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.

2.3 Social Network Analysis

In addition to the profile data, we incorporate the identity features of the top 40 friends within the system, utilizing the nature of social networks. Observing and mining the properties of a social network has the advantage of providing a wider picture about the community. We believe that individual characteristics have a significant influence on the community structure. This leads us to analyze the out-degree, in-degree, overlapping (mutual friends), centrality and isolation of nodes 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 friend. Social network analysis significantly improves the performance of the classifier and in particular, improves the accuracy of the accountability features such as popularity and sociability.

It is well know that social networks are examples of small world networks [15]. A community can be modeled as a network N=I, F, where 'I' represents an individual or node and 'F' represents a friends link or edge. Our data set initially contained over 4.8 million profiles. After removing mutual friends it currently has 2.2 million nodes with 2.4 billion edges between them. This suggests the probability that friend's of friend becoming a friend is much higher than having a stranger as a new friend. The statistical analysis shows that our sample network employs many high degree connections which strongly clustered with an average of 1010 friends for 'public' profiles and 5792 for 'band' profiles.

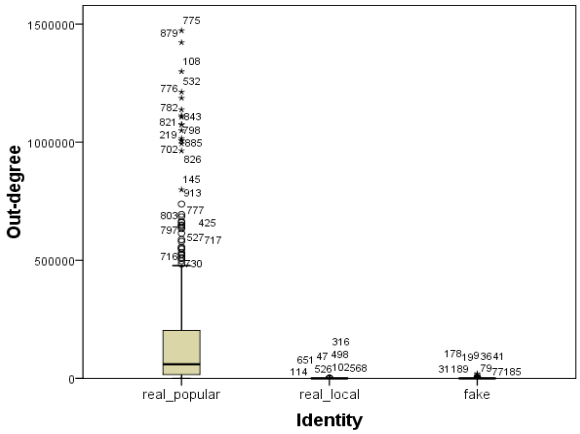


Fig.2. The out-degree distribution according to the type of identity

We can see in Figure.2 the high out-degree distribution of 'real-popular' nodes, while the out-degree analysis alone is unable to verify a 'fake' person from 'real-local'. We use the centrality measurement accumulated from out-degree and in-degree distribution as the 'popularity' factor in our classifier.

Figure.3 illustrates the training nodes (known profiles) position within the network structure. We can see that the 'real-popular' nodes are more closely linked to each other, while 'real-local' nodes are distributed and the majority of nodes with zero out-degree are correlated to 'fake' nodes. Also there is little connection between 'real-popular' and 'fake' users. We can therefore state that social network analysis reveals characteristics about community and the centrality has a high correlation with the type of identity.

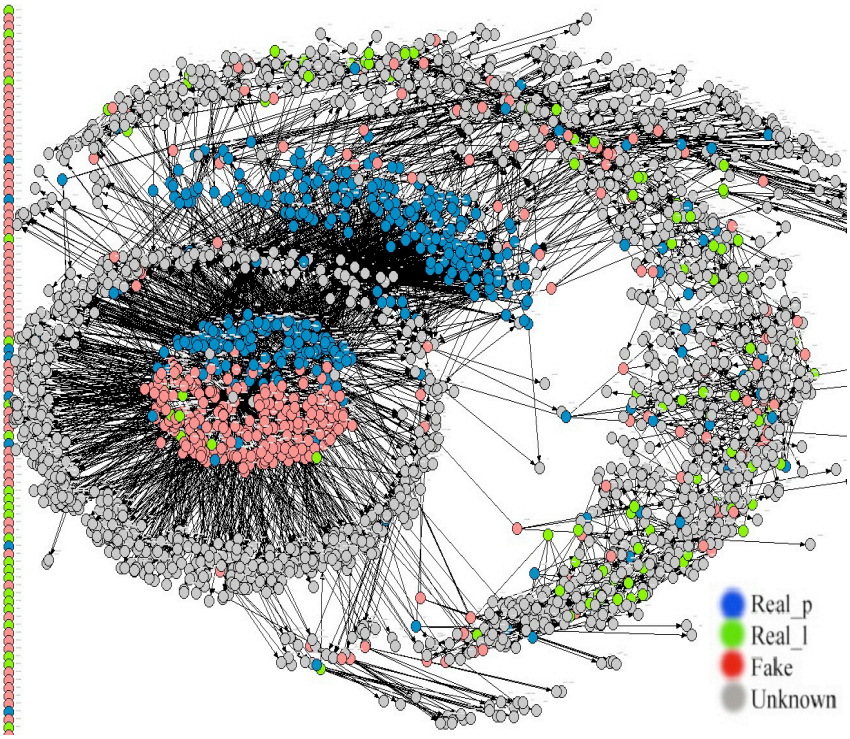


Fig.3. The network structure of known identities

Friendship or linkage also relies on some degree of similarity to imply a connection between people. The similarity measurement reveals more information about the context of link between users and the correlation between their identity types. The following formula used to compare 'I', which represents the value of personality features for individuals with 'F', where implies the average friends' value of personality attributes. The average similarity calculated by dividing minimum value to maximum value between individuals and their friends.

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

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

It is interesting to uncover which identity elements are more important in choosing a friend. For instance, analyzing the pre-classified attributes shows 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' value due to less visibility characteristic of honesty.

2.4 Reality Algoritm

The following formula describes our identity model, where R(x) refers to a 'real' and F(x) to a 'fake' persona and attribute 'a' represents each attribute extracted from profiles. 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, we learned that the choice of friends have influence on the individuals' type of identity. We assumed square root for friends because the friend's attributes are less significant compare to the individual attributes.

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

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

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

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

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

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

Table 2. Acronyms used within preprocessing algorithms

Acronym	Description	Acronym	Description
R(x)	real persona	F(x)	fake persona
H(a)	honest	D(a)	dishonest
A(a)	accountable	U(a)	unaccountable
x	persons	a	attribute
n	number of attribute	f	number of friends

3 Classifier Experiments

Machine learning techniques have been applied in many fields [6,19]. In order to identify patterns within our data and to improve our classifier model in parallel, we conducted supervised learning. We trained multiple classifiers such as basic rules, decision tree, Nave Bayes and nearest neighbor [23] to classify data more efficiently.

Principal Component analysis [24] used to reduce the amount of dimensions required to cluster the data. We used this redundancy technique to examine the correlation between features within the training dataset by generating the main components with minimal loss of information. This analysis indicates which factors or components are more significant when examining personal information to predict the truth about identities. Using both original and pre-classified data, we used two thirds of our known data as training set (to build a model) and one third as test set (to measure the model). There are several validation schemes to estimate the performance of a learner, such as simple validation, regression performance and T-test [23]. We applied the cross-validation operator 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.

In order to achieve more accurate precision we validated different learners for both original data (such as age, gender, location, etc) and pre-classified data (such as valid, popular, traceable etc). We can compare the average performance improvement across all three inputs for the machine learning models used in Table.3. The graph shows that the overall performance over pre-classified data is higher than using the original data, whilst incorporating social network data improves performance yet further. Although the diversity of information in pre-classified data is less, it is much faster and the prediction performance is much effective than using the original data by 84.3% to 66.5%.

Table 3. Average learner performance comparison

	Decision Tree	Nearest Neighbor	Nave Bayes	PCA	Overall Accuracy
Original Data	69.25	67.05	60.99	68.65	**66.49**
Before SNA	81.35	73.09	77.02	81.05	**78.13**
Classifier Performance	86.1	84.59	78.03	88.5	**84.31**

The performance accuracy of the learner produces a confusion matrix [23], which is an evaluation technique to factor a matrix of true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN), 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 described in Table.4 where the overall accuracy can be calculated as: Accuracy= $a+b/a+b+c+d$ where Precision= $a/a+c$ and Recall= $a/a+b$

Table 4. Confusion Matrix

	Actual (real)	**Actual (fake)**
Predicted (real)	TP (a)	FP (b)
Predicted (fake)	FN (c)	TN (d)

Table.5 demonstrates the performance of pre-classified data using 'principal component method. The diagonal line represents the TP and TN prediction with the average performance at 87.1%.

Table 5. Confusion matrix of Principal Component learner over known data

Predicted	Actual Identity			
	real- popular	real -local	fake	class precision
real-popular	372	14	22	91.18%
real-local	11	64	7	78.05%
fake	34	40	428	85.26%
class recall	89.21%	54.24%	93.65%	Accuracy: 87.1%

4 Discussion

We can clarify our assumption which suggests the level of honesty and accountability has a strong correlation on deciding between real and fake personas. As shown in Figure.4 the 'real' nodes are more associated with higher accountability and honesty value, while 'fake' users have less value for both attributes.

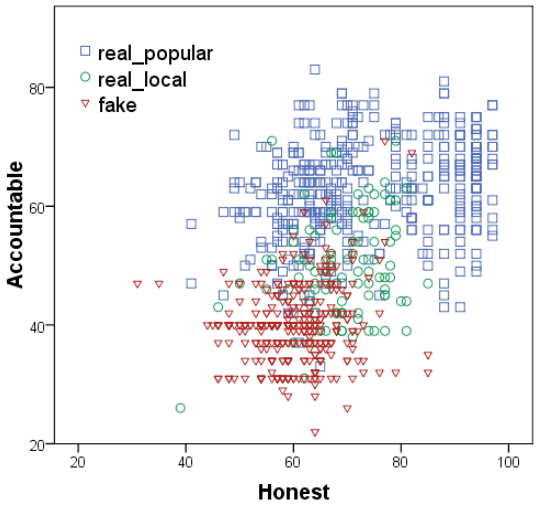
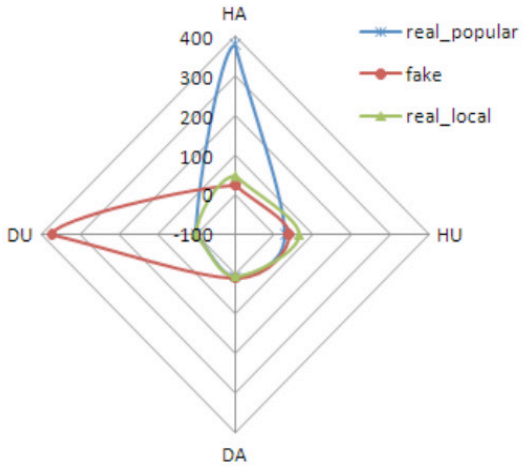


Fig.4. Relationship between honesty and accountability to determine the type of identity

Using our identity model (in Figure.1) we expected four types of identity representation such as honest and accountable (HA); dishonest and unaccountable (DU); honest and unaccountable (HU); dishonest and accountable. Figure.5 illustrates the fraction and frequency of each types of identity representation and proved our assumption that:

- `HA` highly correlated with real-popular users
- `DU` highly correlated with fake users
- `HU` mainly correlated with real-local users
- `DA` mainly correlated with real-local and fake users



`Fig.5.` Four-dimensional identity representation

However, despite the important role of honesty in social interaction it is surprising that the honesty value (quality in content) has less affect than accountability value (quantity in interaction) for identity prediction in MySpace social network.

5 Related Works

Online social networking and identity representation are active research areas with input from computer sciences, statistics, sociology, and psychology.

Studies on psychological aspects of social identity representation examine the social implication of displaying public identities [7]. Toma et al [20] addresses the self-presentation issue by observing the characteristics of users to establish the truth about online dating profiles. Other researchers [9], also used a framework to predict users' identity upon their self-presentation history, while their proposed algorithm achieved high accuracy on prediction, their method is not able to clarify if the predicted traits are real or fabricated.

Many recent studies have shown the efficiency of data mining techniques and analysis of crime detection [19]. Dokas et al employed data mining methods for constructing intrusion detection model to uncover the important patterns in data [6]. Other studies have engaged on detecting a liar online [22], for instance [14]

introduced a mathematical model to investigate the impact of liars; assuming liars either have extremely positive or extremely negative behavior to influence their network. The authors in [3] proposed an automated tool to identify deception in the nonverbal communication to reveal an individual's emotional and cognitive status, such as active/ passive and tense/relaxed behavior.

There are some recent academic studies on online social interaction using popular networks. Carverlee and Webb studied the characteristics of MySpace profiles based on facets of this social network [5]. This paper has similarity to our work, however the focused was only on two identity elements of sociability and use of language within different type of gender. The works on other social network such as Facebook also have focused on identity presentation and information sharing of student network [1,8]. Tufekci [21] also examined the disclosure behavior on MySpace and Facebook users in correlation to privacy issues. The authors in [11] proposed a methodology for clustering and identifying similarity in user's behaviors on YouTube data.

Previous studies on social network analysis, attempted to understand the properties of social network structure [4]. Authors in [13] observed four popular social networks and discussed the correlation within in-degree and out-degree distribution, suggesting that a high degree connection in social network is subject to epidemic. Ryberg and Lasren [16] argue that the notion of exploring weak and strong ties is a valuable contribution to network structure and that the real performance of identity becomes inherently relational. Other works also analyzed social network structure to identify a potential attacks by mining the interaction behavior of the email spam using LiveJournal dataset [17]. The previous and current studies mainly focus on the demographic result, network characterization and privacy implication of social networking. These studies fundamentally differ from our work, as we concentrate on determining the validity of identity on online social community by mining the self-described data.

6 Conclusion

The aim of this study was to examine online identities within the MySpace community to detect real and fabricated users by mining the content of profiles. We proposed a model of personality factors, taking account of the location within the social network and a preprocessor to reduce the dimensionality of the data. We generated a number of classifiers, of which the most effective was a principal component learner, which achieved an accuracy of 84%.

A number of limitations need to be taken into account for this study. Our training set was small, and greater accuracy could be achieved with more data. We have not compared our work against other social networking sites such as Facebook. If a user is really attempting to pass off a fake identity, would the classifier be effective? In future work, we will look at how resilient our classifier is to carefully crafted identities, and how this classifier can be used as evidence in computational trust systems. Despite these limitations, our work provides evidence that it is possible to automatically determine the validity of an online profile.

References

[1] Acquisti, A., Gross, R. (2006). Imagined communities: awareness, information sharing and privacy on the Facebook. Proceeding of the 6th workshop on privacy enhancing technologies.

[2] Boyd, D. (2007). Why youth (heart) social network sites: The role of networked publics in teenage social life. MacArthur foundation series on digital learning - youth, identity, and digital media volume (ed. David Buckingham). Cambridge, MA: MIT Press.

[3] Burgoon, J., Adkins, M., et al. (2005). An approach for intent identification by building on deception detection. HICSS'05, Hawaii.

[4] Casciaro, T. (1998). Seeing things clearly: Social structure, personality, and accuracy in social network perception. Social Networks, 20: 331-351.

[5] Caverlee, J., Webb, S. (2008). A large-scale study of MySpace: Observations and implica-tions for online social networks. In Proc. of ICWSM.

[6] Dokas, P., Ertoz, L., et al. (2002). Data mining for network intrusion detection. Proceed-ings of NSF workshop on Next Generation Data Mining.

[7] Donath, J., Boyd, D. (2004). Public displays of connection. BT Technology Journal, vol 22, No 4.

[8] Ellison, N., Lampe, C., Steinfield, C., (2006). Spatially bounded online social networks and social capital: The role of Facebook. Annual conference of the International Communication Association (ICA).

[9] Hu, J., Zeng, H., et al. (2007). Demographic prediction based on user's browsing behaviour. Proceedings of the 16th international conference on World Wide Web.

[10] Kagal, L., Finin, T., Joshi, A. (2001) Trust-based security in pervasive computing environments. In IEEE Communications.

[11] Maia, M., Almeida, V., Almeida, J. (2008). Identifying user behaviour in online social networks. Proceedings of the 1st workshop on Social Network Systems, ACM.

[12] Mazar, N., Amir, O., Ariely, D. (2007). The dishonesty of honest people: A theory of self-concept maintenance. Journal of marketing research.

[13] 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).

[14] Mundinger, J., Le Boudec, J. (2005). The impact of liars on reputation in social networks. In proceedings of Social Network Analysis: Advances and empirical applications forum.

[15] Newman, M., Watts, D., and Strogatz, S. (2002). Random Graph Models of Social Networks. PNAS, pp 2566-2572.

[16] Ryberg, T., Larsen, M. C. (2008). Networked identities: understanding relationships between strong and weak ties in networked environments. journal of Computer Assisted Learning, 24, 103-105.

[17] Shrivastava, N., Majumder, A., Rastogi, R. (2008). Mining (social) network graphs to detect random link attacks. In proceeding of the 24th ICDE 2008.

[18] Somanathan, E., Rubin, R. (2004). The evolution of honesty. Journal of Economic Behavior and Organization, 54, 1-17

[19] Thongtae, P., Srisuk, S. (2008). An analysis of data mining applications in crime domain. CIT Workshops 2008.

[20] Toma, C. L., Hancock, J. T., Ellision, N. B. (2008). Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles.

[21] Tufekci, Z. (2008). Can you see me now? Audience and disclosure regulations in online social network sites. Bulletin of Science Technology and Society, vol. 28, no. 1, pp.20-36.

[22] Whitty, M. T. (2002). Liar, liar! An examination of how open, supportive and honest people are in chat rooms. Computers in Human Behavior, 18, 343-352.

[23] Witten, I. H. and Frank, E. (2000). Practical machine learning: Tools and techniques with Java implementations, Morgan Kaufmann, San Francisco.

[24] Zou, H., Hastie, T., Tibshirani, R. (2004). Sparse principal component analysis. Journal of Computational and Graphical Statistics.