2014

IDENTIFYING FAKE PROFILES IN LINKEDIN

Shalinda Adikari
Department of Information Systems, National University of Singapore, Singapore, shalinda.adikari@gmail.com

Kaushik Dutta
National University of Singapore, duttak@nus.edu.sg

IDENTIFYING FAKE PROFILES IN LINKEDIN

Shalinda Adikari, Department of Information Systems, National University of Singapore, Singapore, shalinda.adikari@gmail.com

Kaushik Dutta, Department of Information Systems, National University of Singapore, Singapore, duttak@nus.edu.sg

Abstract

Social networks have become an everyday tool in our lives and different social networks have different target groups. Among them LinkedIn is greatly preferred by the people who are in the professional occupations. With the rapid growth of social networks, people tend to misuse them for unethical and illegal conducts. Creation of a fake profile becomes such adversary effect which is difficult to identify without apt research. The current solutions that have been practically developed and theorized to solve this contention, primarily considered the characteristics and the social network ties of the user's social profile. However, when it comes to LinkedIn such behavioral observations are highly restrictive in publicly available profile data for the users by the privacy policies. The limited publicly available profile data of LinkedIn makes it ineligible in applying the existing approaches in fake profile identification. Therefore, there is a need to conduct targeted research on identifying approaches for fake profile identification in LinkedIn. In this research, we identify the minimal set of profile data that are necessary for identifying fake profiles in LinkedIn and identify the appropriate data mining approach for such task. We demonstrate that with limited profile data our approach can identify the fake profile with 84% accuracy and only 2.44% false negative, which is comparable to the results obtained by other existing approaches based on the larger data set and more profile information.

Keywords: LinkedIn, Fake Profile, Neural Network, Support Vector Machine, Principle Component Analysis, Data mining.

1 INTRODUCTION

In recent past, social networks have made a drastic change in the social life and it changed the web into "social web" where users and their communities are the centres for online growth, commerce, and information sharing (Rheingold 2000). Social networks have a unique value chain which targets different user segments. To find an old friend, we used to peruse Facebook, but if it is to access micro blogging then we have Twitter. LinkedIn is well known to maintain a professional resume with a high degree of contacts or to find a contact from a professional group. Among these Facebook becomes the most visited social network and it has 800 million visitors per month while Twitter becomes the second best social network with 250 million visitors per month (eBizMBA 2014). In recent years LinkedIn becomes the next evolutionary in social networks, as the world's largest professional network. It has more than 200 million visitors per month (eBizMBA 2014).

This surge of social networks' popularity and the availability of large amount of information from users' email addresses to their personal messages make them easy targets to the adversaries. Most of these targets focus on retrieving user information without user consent. For that, by intruding into the user profile or connecting with the user through fake profiles are considered as the mostly practised techniques (Fire et al. 2012). With the advancement of social network security it becomes tremendously difficult to infringing into social networks. Resultantly, now adversaries create fake profiles to get the access to other accounts. As per the statistics of Cloudmark, around 20-40% of the Facebook accounts could be fake profiles and this is fairly similar with Twitter (Lee et al. 2011). Due to high amount of user involvement and millions of daily transactions it becomes hard to detect suspicious user behaviors in the network and separate them from the legitimate users. Conversely, effort is taken to find those malicious accounts and flag them as fake, yet it did not achieve the results as expected. This becomes more complex, attributable to, constricted user privacy policies, restrictions for data collections and difficult to distinguish between the fake and legitimate profiles. Even though most of the preceding research targeted on identifying profile cloning, spam information distribution, and intrusion detection (Kontaxis et al. 2011; Fire et al. 2012), now it becomes the time to draw extra attention to finding solutions to differentiate legitimate and fake profiles in a sensible manner.

1.1 Background

A fake profile is a social network profile of a person who maintains a false identity in the internet to pretend as someone else. It is found out by (Krombholz et al. 2012) the fake user behavior is different from the legitimate users. Therefore, the amount and the type of information that a fake user pass into their profile have a clear discrepancy from the legitimate user. Among the several ways of creating fake profiles following three ways can be considered as the significant. First is fabricating the own profile (Krombholz et al. 2012) which is used to increase the discernability of niche content and manipulate the attraction towards the profile (Cao et al. 2012; Bilge et al. 2009). Next method is profile cloning (Jin et al. 2011; Kontaxis et al. 2011) where offender creates a similar profile of the legitimate user in the same or another social network by copying the victim's profile and adding victim's friends into new fake profile. The last method is creating a profile with a fake identity (Krombholz et al. 2012). The trick on such profiles is perpetrators first attain the victims trust and confidence, and then cheat on them by collecting the confidential information. Due to the similarity of the features of legitimate and such fake identities, it is immensely difficult to distinguish them without ascertaining reliable data.

LinkedIn is considered as the highly recommended social network website for professionals and it has an unprecedented growth compared to other similar social networks. It is a renowned digital resume where users can list all their bio data and keep list of professional contacts. In addition, there is a high trend of recruiting people for different professions referring to the LinkedIn profiles. Indeed, LinkedIn

is a digital human resource directory where recruitment agents can select potential workers or users can contact such available options. Attributable to all these facts, creating fake LinkedIn profiles is increased extremely. One such latest scenario[1] was, in the beginning of 2014 some hackers executed a Botnet attack and created thousands of fake LinkedIn profiles. The present process of identifying fake profiles in LinkedIn is limited to manual reporting of such profiles. When a LinkedIn user suspects a particular profile to be faked, he or she can use the fake profile flagging option to notify the LinkedIn team about his or her apprehensiveness.

1.2 Problem

The focus of this research is recognizing and differentiating the legitimate profile and fake profiles in LinkedIn. Most of the solutions developed to address the above issue is based on either Facebook or Twitter. These social networks have rich and fully functional Application Programming Interfaces (API) to acquire relevant, real-time and up-to-date user information in analogous to the research requirements. Facebook API[2] facilitates to access profile information like user activities, friends' activities, friends of friends and most of the basic user details (age, birthday, profile status, relationship status, likes, group details etc.). Similarly, Twitter API[3] provides twitter counts, followers, notifications, friends, basic user details etc.

The profile data in social network consist of two main parts, static and dynamic. Former is about the information which is set by the user statically, while the latter is observed by the system and is the result of user's activity on the social network. The static data typically includes users' demographics and interests, and dynamic data relates to user activities and position in the social network (Kazienko and Musiał 2006). Most of the existing research solutions depend on both static and dynamic data, which is inapplicable for LinkedIn, where it has merely a less number of visible static profiles and no dynamic profile details to the public. Due to its privacy policies and very restricted information visibility (Bradbury 2011), none of the existing practical and theoretical means of fake profile detections are feasible to apply. Therefore, in this research our goal is to identify an approach to determine legitimate profiles and fake profiles in LinkedIn.

1.3 Challenge

The key challenges that we come across doing this research are data collection and fake profiles identification of LinkedIn. Due to LinkedIn privacy policies, gathering data from it is highly restricted. We can access very limited profile characteristics and number of profiles via its API. Even to access certain basic LinkedIn user information such as education details, date of birth, suggestions, telephone number, total number of connections/skills and expertise we need the user permission. In addition, through the LinkedIn API we are offered only to access first degree level user information yet with the normal web user interface, it provides access up to the third degree level of connections' information.

Secondly the access to actual fake profiles in LinkedIn context is greatly unattainable. However, we were able to find a list of web sources where certain LinkedIn fake profiles have been manually identified and listed. So, unlike previous research, in this research we have used only the authentic fake profile data for the research instead of simulating the fake profiles.

1.4 Methodology

In this research, we considered four data mining techniques, Neural network (NN), Support vector machine (SVM) and Principal Component Analysis (PCA). All these are well known and commonly

[1] Source: http://dailyglobe.com/38505/linkedin-sues-unknown-hackers/
[2] Source: https://developers.facebook.com/docs/reference/fql/user/
[3] Source: https://dev.twitter.com/docs/api/1/get/users/show

used data mining techniques. In much social network research, neural network and SVM are adopted as the principle mining techniques. Few such research areas are spam message identification, profile cloning and intruder detection. PCA is applied to reduce the number of dimensions of the datasets (Jolliffe 2005) .

In summary, we have developed a technique for identifying the approaches to identify fake profiles in LinkedIn by combining multiple data mining techniques. We have compared and discovered the appropriate data mining technique to identify fake profiles in LinkedIn with minimal amounts of profile data. We have demonstrated that our approach performs with an accuracy of 84% and false negative of 2.44%, which is comparable to the results reported by existing research – that is based on other social network data and much in-depth profile data.

The rest of the paper is organized as follows. Section two provides an overview of the research carried out in related to LinkedIn network and prior research on fake profile identification. In section three, we describe the LinkedIn dataset and the mechanism followed to collect data. Section four explains the methods used in the construction and evaluation of the each technique and their results. In section five we discuss the overall comparison of the accuracy rates. Section six identifies limitation of the study and future directions and finally, section seven we present our conclusions from this study.

2 RELATED WORK

In this section we provide some insight into the existing research on the LinkedIn network with an example of cloning attack identification. Moreover, we describe existing work carried out in related to fake profile identification and in similar research background.

2.1 Research based on LinkedIn data

So far, little research has been carried out accounting LinkedIn as the primary data source. Hsieh et al. (2013) have conducted a research on LinkedIn to understand the probability of connections between two people based on their organizational overlap. Xiang et al. (2010) have used interaction activity and similarity of user profiles to develop an unsupervised model to estimate friendship strength. They evaluated the system on proprietary data from LinkedIn.

2.2 Determine fake profile

In the midst of the different strategies that have been developed to determine fake profile in social networks, many of them follow the similar portfolio of techniques, but they have been applied in different contexts (in different social networks or on different feature set). Here we discuss only the selected unique solutions related to LinkedIn and other social networks.

2.2.1 In LinkedIn

Kontaxis et al. (2011) has developed an approach of detecting cloned profile based on LinkedIn data. The approach consists of three components - distiller, profile hunter and profile verifier. The information distiller constructs test queries using the information extracted from the profile and run them in search engines and social networks. Then results returned against each query are taken into account by the distiller to create the user-record. The user record is a set of user identifying terms along with the user's full profile name. This record is the input to the next component. The output of the information distiller is used by profile hunter to locate potential social network profiles belonging to the user. All the returned results are grouped as a profile-record. The profile record contains a link to the user's real profile and to all other returned profiles. The next component, profile verifier, examines the profile-record for similarity check with the user's original profile. Through the profile verification a similarity score is calculated based on the common values of information fields. Finally the profiles which have high probability to be cloned are presented with similarity scores.

2.2.2 In other social networks

Fire et al. (2012) used topology anomalies to identify the spammers and fake profiles. Apart from domain of graph theory and supervised learning, they exerted the parallel decision tree and Naïve Bayes classifiers into their algorithm. Boshmaf et al. (2011) adopted the traditional web based Botnet design to build a group of adaptive social-bots as a socialbot network and analyzed its impact via millions of the Facebook users. Jin et al. (2011) analyzed the behavior of identity clone attacks and proposed a detection framework. Cao et al. (2012) ranked users in online services to detect fake accounts. Their ranking algorithm is supported by social graphs according to the degree-normalized probability of a short random walk which resides in non-Sybil region. As a case study Krombholz et al. (2012) have analyzed privacy related issues in social media contexts by creating desirable fake Facebook profiles and interacting with existing legitimate users of the network. Consequently they could discover how much information that can be harvested and analyzed from the users who interact with these fake profiles

3 LINKEDIN DATASET

As stated in the beginning of the paper LinkedIn is regarded as the primary data source for our research. Due to existing privacy policies and system API limitations of LinkedIn, we could get hold of few profile features only. To collect real fake profiles, we browsed the web and found several blogs and web sites where people have identified and listed the fake profiles. In some cases same profiles have been recognized by different people as fake profiles. Conversely, profiles which are specified as fake by only an individual is re-confirmed by manually checking whether they are fake. In the manual process of fake profile detection we followed the most commonly considered techniques by the social network community, such as, groups details are not matching with the users' other profile data, connected to other fake profiles, the information are not logical or reasonable, profile data is disharmonized and recommendations are made only among fake profiles, there are no credible connections in the connection chain and checking the legitimacy of the profile picture by searching on Google and TinEye[4]. For example, Figure 1 shows two sample fake profiles. Through this process, we were able to confirm 34 fake profiles. Next, we have randomly identified 40 legitimate profiles from the LinkedIn public profiles and confirmed their legitimacy through aforementioned manual techniques.

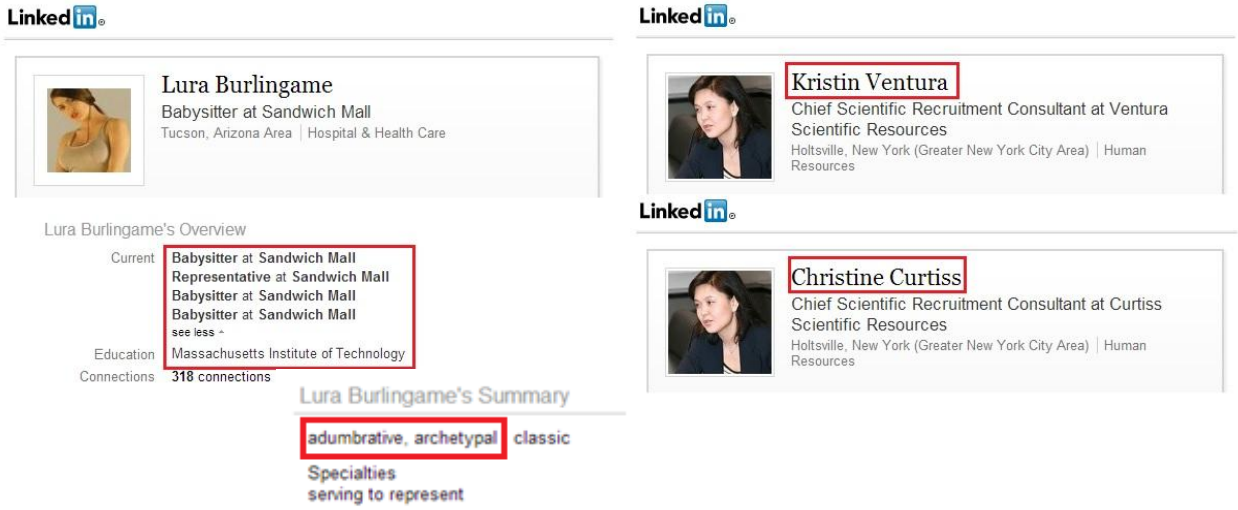

Figure 1. Examples for LinkedIn fake profiles: (1)Bot generated content (left side profile)[5], (2)two profiles with the same content and different names (right side profile)[6]

[4] Source: http://www.tineye.com/
[5] Source: http://www.slideshare.net/augustinefou/fake-profiles-on-linkedin
[6] Source: http://www.dikomci.com/post/37712291401/how-to-spot-a-fake-profile-on-linkedin-and-facebook

Table 1 lists all the profile features which we were able to capture publicly in the both legitimate and fake profiles along with the maximum and average values of each profile feature across all profiles in our dataset.

Due to the LinkedIn restrictions, in variant of the actual value, the highest number of connections is 500 and the highest number of skills is 50. Therefore, rather than computing the normalized values via mean and standard deviation we utilized the maximum and minimum value of each feature. As each feature value doesn't present at least once in either a legitimate or a fake profile, the minimum value for all the features is 0. The maximum and average values are shown in Table 1.

Profile feature	**Maximum value**	**Average value**	**Description**
No_Languages	5	0.347	Number of languages can speak
Profile_Summary	1	0.52	Presence of profile summary
No_Edu_Qualification	7	1.467	Number of education qualifications
No_Connections	500	294.867	Number of connections
No_Recommendation	37	2	Number of recommendations made
Web_Site_URL	1	0.28	Presence of a URL for personal website
No_Skills	50	10.213	Number of skills and expertise
No_Professions	16	3.08	Number of past and present professions
Profile_Image	1	0.76	Presence of a profile image
No_Awards	10	0.56	Number of awards won
Interests	1	0.267	Presence of any type of interests
No_LinkedIn_Groups	51	8.907	Number of LinkedIn groups and association added
No_Publications	16	0.613	Number of publications made
No_Projects	7	0.24	Number of projects that work
No_Certificates	9	0.267	Number of certificates hold

Table 1. Details of the profile features

Finally, we divided all profiles into two equal groups randomly by three times (See Table 2) such that each of them contains the same number of fake and legitimate profiles. Moreover, each profile is distinctive from one another. Thus we have 3 data sets (Dataset 1, Dataset 2, Dataset 3), where each set contains two groups of 37 numbers of profiles – one of the group is marked as Training dataset and the other is marked as Test dataset. Each of these groups of 37 profiles has 20 legitimate profiles and 17 fake profiles.

	Training dataset	Test dataset
Legitimacy Profiles	20	20
Fake Profiles	17	17

Table 2. The number of fake and legitimate profiles used in training and test data sets

4 METHODS AND RESULTS

In this section we explain how each technique is used in the process of data mining to differentiate legitimate and fake profiles. The process has three levels, in the first level profile features are extracted by PCA and then in the second level NN and SVM are used to determine the fake and legitimate profiles. The third level, we calculate and compare the accuracy rates across the results of both techniques (see Figure 2).

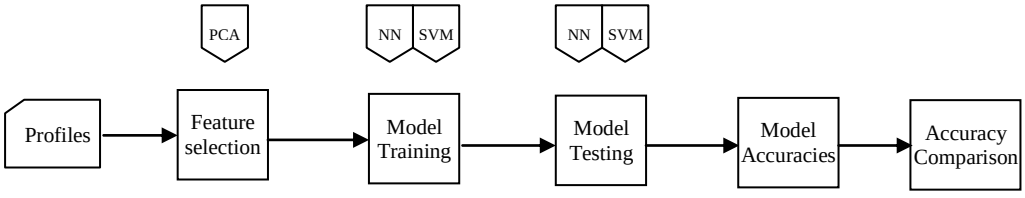

Figure 2. Design approach to calculate accuracy rates for data mining techniques

4.1 Principal component analysis

In this research PCA plays a major role by providing the support to take the decision on which profile features to be used in the data mining. PCA is considered as the simplest but robust dimensionality reduction technique.

Among the number of different mathematical ways of deriving PCA results we have selected the simplest case that is the variance maximization. In variance maximization first principal component has the highest projection variance which is the direction in feature space along and the second component defines the direction which has highest projection variance among all the other orthogonal direction to the first component. In the process of calculating the score on the features, both fake and legitimate profiles are considered. We have used eigendcomposition which is the most commonly practiced calculation methodology for PCA to find the number of components. Initially, to ensure the sampling adequacy we have tested for Kaiser–Meyer–Olkin (KMO) and Bartlett's Test. The resulted KMO value is 0.724 which is higher than the acceptable level of 0.5 and the Barlett's test is significant at $p<0.05$ (Ashcroft and Parker 2009). This verifies the required numbers of samples to be adequate to precede the study.

Then we estimated the variation of the components and selected the components which have Eigenvalues more than 1 (Kaiser 1974) . Eigenvalues provide information of the variability in the data. There we found 5 components which have the total variance of 64.82%. Each component variations and their Eigenvalues are demonstrated in Table 3.

	Initial Eigenvalues		
Component	**Total**	**% of Variance**	**Cumulative %**
1	4.375	29.170	29.170
2	1.801	12.008	41.177
3	1.290	8.601	49.778
4	1.195	7.970	57.748
5	1.061	7.075	64.823
6	.934	6.225	71.049
7	.840	5.603	76.652
8	.820	5.464	82.116
9	.618	4.119	86.234
10	.522	3.479	89.714
11	.440	2.936	92.650
12	.369	2.462	95.111
13	.284	1.891	97.002
14	.242	1.611	98.613
15	.208	1.387	100.000

Table 3. Total Variance Explained by PCA

Then we checked each component feature score which provides information about the structure of the observations and identified features that either load into several components with the score value of more than 0.5 or not load into any component with more than 0.5 score value (Kaiser 1974). For these features to better understand the relationship between features and extracted principal components, we used Varimax rotation to load the features into the components again. Still, we found some features unintendedly load to several components without acceptable score values. To get clear feature loadings for the components we have removed such features step by step as mentioned in the following algorithm. Finally, we could obtain the results as depicted in Table 4 by removing *Profile_Image*, *No_Awards*, *No_LinkedIn_Groups* and *No_Publications*. Due to the removal of the features, remains are loaded into 4 components with the total variation of 66.15% and even the KMO value reduces to 0.655 which is still higher than the recommended boundary (>0.5) with the same significance value. The detail algorithm of PCA based feature selection is given below. The Table 4

shows the selected features and how they load the principal components at the end of running the feature selection algorithm.

Algorithm for feature reduction through PCA

Initialize F with all features, where f is a feature in feature set F
Initialize each X_f to zero, where X_f is an indicator variable associated to feature f
Do
    Run *PCA* with *Varimax* rotation
    If (*Eigenvalue* ≥ 1)
        Select C, where C is selected principal components
        Initialize L to empty, where L is a list
        For each $f \in F$
            For each $c \in C$
                If $S_f > 0.5$, where S_f is feature scores for feature f
                $X_f = X_f + 1$
            End For
            If X_f is not equal to 1
                Add f to L
            End If
        End For
    End If
    For Each $f \in L$
        Remove f from F
While (L *length* > 0)
Output : F is the set of selected feature set

Feature No.	Profile Feature	Component 1	Component 2	Component 3	Component 4
[1]	No_Languages	0.614	0.098	0.218	-0.162
[2]	Profile_Summary	0.623	0.016	0.247	-0.097
[3]	No_Edu_Qualifications	0.827	0.139	-0157	0.266
[4]	No_Professions	0.702	0.171	0.153	0.311
[5]	Web_Site_URL	0.195	0.860	0.106	-0.046
[6]	Interests	0.079	0.913	0.025	0.040
[7]	No_Connections	0.208	-.0177	0.684	0.157
[8]	No_Recommendations	-0.007	0.161	0.800	0.076
[9]	No_Skills	0.426	0.335	0.695	0.122
[10]	No_Projects	0.164	0.072	0.154	0.685
[11]	No_Certificates	-0.075	-0.082	0.077	0.843

Table 4. Selected feature loadings of PCA

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
[1]	1.000										
[2]	0.261	1.000									
[3]	0.311	0.379	1.000								
[4]	0.287	0.296	**0.624**	1.000							
[5]	0.161	0.236	0.259	0.237	1.000						
[6]	0.155	0.061	0.177	0.266	**0.659**	1.000					
[7]	0.095	0.169	0.105	0.340	0.036	-0.053	1.000				
[8]	0.120	0.251	0.021	0.205	0.161	0.141	0.278	1.000			
[9]	0.502	0.317	0.319	0.448	0.406	0.331	0.465	0.516	1.000		
[10]	0.132	0.153	0.206	0.207	0.075	0.091	0.188	0.132	0.270	1.000	
[11]	-0.065	-0.034	0.123	0.156	-0.090	-0.054	0.101	0.136	0.120	0.319	1.000

Table 5. Selected feature correlation matrix

As per the Table 5 we can see that all the correlations are less than 0.6, only in two combinations (feature [3] and [4], feature [5] and [6]) the values are marginally higher than 0.6. Nevertheless, those features are loading the same component. For example, both feature [3] and [4] are loading the component 1 (Table 4). Similarly feature [5] and [6] are loading component 2 (Table 4). Therefore we can state that the selected features are not highly correlated to each another (Taylor 1990; Shu and Chuang 2011). As a conclusion of this PCA based feature selection step we summarize the selected features in Table 6.

Feature Name	**Is Selected by PCA? (Y – Yes, N – No)**	**Selected Feature Number**
No_Languages	Y	[1]
Profile_Summary	Y	[2]
No_Edu_Qualification	Y	[3]
No_Connections	Y	[4]
No_Recommendation	Y	[5]
Web_Site_URL	Y	[6]
No_Skills	Y	[7]
No_Professions	Y	[8]
Profile_Image	N	
No_Awards	N	
Interests	Y	[9]
No_LinkedIn_Groups	N	
No_Publications	N	
No_Projects	Y	[10]
No_Certificates	Y	[11]

Table 6. List of Selected and all profile features

4.2 Neural network

Currently there are many neural network (NN) algorithms that are used to train models either through supervised learning or unsupervised learning. In this research our focus is on the supervised learning where we have the legitimacy as the response variable and selected profile features (Table 6) as the input. We selected the Resilient backpropagation (Rprop) algorithm as the base algorithm. Rprop does not account for the magnitude of the partial derivatives (only the sign) of the patterns and work out independently on each weight (Riedmiller and Braun 1992). Rprop is considered as one of the fastest algorithm in data mining (Kumar and Zhang 2006). We selected *neuralnet* package in R project for statistical computing (Günther and Fritsch 2010). *neuralnet* is flexible to include the custom-choice of error-function, number of covariates with response variables and the number of hidden layers with hidden neurons.

Since response variable (legitimacy of the profile) is considered as binary (if legitimate, then the value is '1' and if fake, then value is '0'), logistic function (default) is chosen as the activation function of the training and cross-entropy (err.fct="ce") is selected as the error function. To ensure that the output is mapped by the activation function to the interval [0, 1]; we defined linear.output as FALSE (Günther and Fritsch 2010). With this preparation, we trained the model by determining the number of hidden neurons and layers in relation to the optimized results. After several iterations, the best result (i.e. highest accuracy) is achieved with one hidden layer with two neurons.

First, we trained the model for all three datasets with all the features and saved their models in different variables. Then again we selected same datasets, remove the features which are not selected by the PCA and saved the models to different variables.

The "compute" function of the library is used to predict results for new data based on the stored NN models. Since the compute function automatically redefine the NN structure only to calculate the output for arbitrary covariates, we could easily figure out the predictions for the legitimacy of each related test datasets with all and selected features. Then the results are compared with the actual

legitimate values (i.e. whether the profile is fake or legitimate) and calculated the accuracy for each dataset with all and selected features in Table 7.

Here we have defined the accuracy as

$$\%\, Accuracy = \frac{Total\ number\ of\ correctly\ identified\ profiles, both\ fake\ or\ legitimate}{Total\ number\ of\ profiles} \times 100$$

We can see from Table 7, that the accuracy result is higher in the case of selected features than when all features are used in the NN. In case of all features, the model detoriates due to unnecessary data points leading to the over-fitting problem of NN. This definitely shows the importance of the PCA step in our approach if NN is used for detection of legitimacy of a LinkedIn profile.

	Dataset	Training error	Accuracy (%)
All features	Dataset 1	0.043	84.85
	Dataset 2	0.083	68.29
	Dataset 3	0.064	86.11
Selected Features	Dataset 1	0.025	87.88
	Dataset 2	0.089	70.73
	Dataset 3	0.012	89.89

Table 7. Accuracy of the results obtained through neural network training

4.3 Support vector machine

In this section we apply Support Vector Machine (SVM) based approach to identify the fake profile. For the SVM training we applied C-support vector classification (C-svc) which is a Quadratical Programming (QP). C-svc can find the best possible hyperplane by measuring the margin between two classes using 2-norm of the normal vector and norm-1 is used for the feature selection (Zhang et al. 2013) According to the Mercer's theorem (Cortes and Vapnik 1995) the kernel function K can be considered as equal to a dot product in input space and due to the nonlinearity of the profile features, SVM is able to create a random decision functions in the input space on the kernel function.

Both the Radial Basis function kernel (rdfdot) and Polynomial kernel (polydot) are used as kernel functions for better understanding of SVM performance on the dataset.

The Radial Basis kernel is, selected because it uses the heuristic in sigest to calculate better sigma value, and we did not need to assign values to the kernel parameters. Radial basis function kernel K can be written as

$$K(X_i, X_j) = \exp(\gamma |X_i - X_j|^2)$$

Polynomial kernel is selected as it uses a combination of features of the input sample instead of determining similarity of those independently. The polynomial kernel function can be written as,

$$K(X_i, X_j) = (\gamma\, X_i \bullet X_j + C)^d \qquad \text{When C=0 kernel is called homogenous.}$$

For both the kernels $K(X_i, X_j) = \phi(X_i) \bullet \phi(X_j)$; $\gamma = -\frac{1}{2\sigma^2}$

The transformation function ϕ maps a dot product of input data points into higher dimensional feature space where the non-linear patterns would demonstrate linearity. γ is an adjustable parameter and $\gamma > 0$

Since we intended to use C-svc classifier, we use KSVM (function of R, kernlab package) to train the SVM model. KSVM facilitates the Sequential Minimal Optimization (SMO) algorithm for solving the SVM quadratic programming (QP) optimization problem (Joachims 1999) We have performed the

training with the two proposed kernel functions (Radial Basis and Polynomial) to create SVM models for all three training datasets with all features and PCA based selected features. Then the models are tested using the test dataset of the respective group. In this way we have total 12 models (2 Kernel functions, 3 datasets – each with both all features and selected features) to test and compare. Each model is tested with the pertinent test dataset and we calculated the accuracy rate. The consolidated results are presented in the Table 8.

	Kernel Type / Dataset	**Radial Basis kernel (%)**	**Polynomial kernel (%)**
All features	Dataset 1	78.79	84.85
	Dataset 2	73.17	73.17
	Dataset 3	88.89	91.67
Selected features	Dataset 1	75.76	84.85
	Dataset 2	78.05	75.61
	Dataset 3	91.67	91.67

Table 8. The accuracy rates of the results based on Kernel and the dataset

In each of the scenarios Polynomial kernel derived the optimized results with less number of vectors in comparison to the Radial Basis Kernel. Since we need to compute the dot product of each support vector with the test point, the computational complexity of the model is linear to the number of support vectors.

We observe from Table 8 that other than Dataset 2 with selected features, Polynomial Kernel performs better or equal to the Radial Basis kernel. Also, the Polynomial Kernel performs better when it is applied on selected features (by PCA) than when all features are considered.

Feature selection		**Radial Basis kernel (%)**	**Polynomial kernel (%)**
All features	False positive	14.84	9.84
	False negative	6.73	6.93
Selected features	False positive	10.94	13.52
	False negative	7.24	2.44

Table 9. False positive and false negative values based on the Kernel and feature selection

Additionally, in Table 9, we present average false negative and false positive values across all three datasets for both the kernels. We can see from Table 9 that the false negative value of Polynomial kernel with selected features is the lowest. It is important to note that in this case false negative has higher risk value in business than false positives. For example, due to false identification of a fake LinkedIn profile as legitimate profile, Human Resource management of an organization may spend unnecessary recruitment cycle costing the organization.

Thus the above discussion concludes for identification of legitimacy of LinkedIn profile with SVM, SVM with polynomial kernel applied on PCA based selected profile features gives the highest accuracy with the lowest false negative.

5 DISCUSSION

In this research, we have compared the results of two data mining techniques to determine the most appropriate approach to differentiate the legitimate profiles from fake profiles in LinkedIn. Table 12 summarizes the final accuracy values akin to each technique by calculating the averages across all three datasets. In addition, it shows the average false positive rate and false negative rate for each technique.

Although the RBF kernel is the mostly used kernel in the data mining context, in our scenario Polynomial kernel gives us the higher accuracy compared to RBF kernel. Additionally Polynomial

kernel false negative value is reduced when PCA based selected features are used (see Table 9), therefore we can conclude in case of SVM, polynomial kernel with selected feature is the right choice.

	Feature selection	Accuracy Rate (%)	False Positive (%)	False Negative (%)
Neural Network	All features	79.75	15.17	5.08
	Selected features	82.83	13.21	4.29
Support Vector Machine (Polynomial Kernel)	All features	83.23	9.84	6.93
	Selected features	84.04	13.52	2.44

Table 10. Accuracy comparison of the two techniques NN &SVM

In accordance to the final accuracy rates, we can see that SVM has the highest accuracy rate between the two techniques regardless of the number of features used. However the difference between NN and SVM is 2.48% when all features are selected and 1.21% when only the extracted features are selected. As per the theoretical rationale SVM vector machine is more preferred data mining technique for the dataset like this, because SVM can compute results even with less number of training data points and it does not suffer from local extrema.

False positive and false negative columns exhibits the percentages of the number of legitimate profiles detected as fake and number of fake profiles detected as legitimate respectively. Compare to the false positive, false negative has a higher risk, because if a fake profile is identified as legitimate, then the impairment can be occurred is much higher whilst a legitimate profile detected as fake. As shown in Table 10, SVM with selected feature has the lowest false negative value (2.44%). Thus, between the two approaches (NN and SVM), *SVM with polynomial kernel gives the most accurate result with low false negative* for the task of identification of fake profile in LinkedIn.

NN and SVM provide higher accuracy when the features are selected through the PCA. For both dataset 1 and 2 the accuracy values with the selected features are higher than when all the features are selected. In addition the false negative value is less for both techniques when only the selected features are used for legitimate determination. Thus, PCA based feature selection is an important step in the process of identification of fake profiles in LinkedIn.

So, from the above discussion, we conclude *PCA based feature selection and subsequently SVM with Polynomial Kernel based modelling for determining legitimacy of profile is the right approach for identification of fake profile from LinkedIn, where limited number of profile features are public.*

SVM accuracy can be advanced by further analysing the kernel, fine-tuning the kernel parameters and tolerance level (Cristianini and Shawe-Taylor 2000). NN is more accurate when there are higher number of data points, thus we can expect more optimized results while the numbers of profiles are increased. In the LinkedIn it is quite difficult to increase the number of data points as LinkedIn impose limitation on accessing its data and it is particularly challenging to increase the number of fake profiles.

Next we show how our result compares with the results of previously proposed approaches. It is difficult to implement and run previous approaches on our dataset, because neither of the previous approach is based on limited LinkedIn data. So, in Table 11, we present the accuracy of the results of previous research as reported by them along with the social media on which it was applied and the data set requirement of the approach.

In summary, prior research which focused on fake profile identification, has similar accuracy rate compared to what we accomplished in our research. In all these prior research, researchers have used the user activities as a criterion to decide the legitimacy of a profile. A user activity of a profile includes all the dynamic information of a user (number of posts, information about friends and their behavior). Such dynamic data of a user are impossible to access in LinkedIn, due to LinkedIn's data accessibility restriction which is elucidated in Section 1.3. Though, the prior studies listed in Table 13 have analyzed more than thousands of profiles to accomplish the shown accuracy, all the fake profiles exploited are simulated. On the contrary, our approach considered actual fake profiles in LinkedIn. Additionally, with the consideration of practicality of the approach, our approach is based on limited

static profile data and does not include any profile data that is hard to access or mostly restricted by LinkedIn profiles. *Considering these significant differences, compared to results of prior research our results of 84% accuracy with 2.44% false negative can be considered as an excellent improvement.*

Technique used	Accuracy (%)	Feature types	Social network	Source
Support Vector Machine	78	Dynamic and Static e.g.: profile age, presence of profile image, followers and friends count, posts/messages, details of tweets	Twitter	(Chakraborty et al.).
Naïve Bayes	67			
Decision Tree	69.25	Static e.g: profile's content such as age, gender, location	MySpace	(Feizy et al. 2009)
Nearest Neighborhood	67.05			
Decision Tree	86.10	Dynamic e.g.: profile's connectivity, the amounts and types of interactions		
Nearest Neighborhood	84.59			
Weka Classifier : Random Forest algorithm	94.5	Dynamic e.g.: number of friends, friend requests, details of short text messages	Twitter	(Stringhini et al. 2010)
	97	Dynamic e.g.: notifications, private message, wall posts, and status updates	Facebook	

Table 11. Details of prior researches on fake profile identification

6 LIMITATIONS AND FUTURE WORK

The main limitation that we see is the verification of the sources and the published fake profiles. There can be situations where the source classifies a profile as a fake profile without proper evidence. Second, when a cloning attack occurs on certain profile we cannot actually identify which profile is the fake. One similar setup is shown in Figure 1. Between two of these profiles one can be legitimate. Our future intention of this study is to follow a similar approach and analyze other social networks to check the status of the accuracy level of differentiating fake and legitimate profiles exclusively based on the limited factual data. Also, we can improvise the data mining by considering other important information such as characters of user name, including length, lower case, and so on; location information including size of address and geographical connection.

7 CONCLUSION

In this paper, we propose an approach to identify the fake profile in LinkedIn with limited profile data. As we concluded in our discussion SVM with Polynomial Kernel on PCA based selected features has the capability to train a model to achieve higher accuracy with low false negative on differentiating the legitimate profiles and fake profiles in LinkedIn.

Many of the past research on fake profile are based on where both dynamic and static behavioral data on social network and in most cases tested only on the simulated fake dataset. Even though there is a research conducted on LinkedIn data in related to spam detection (Prieto et al. 2013), to our knowledge this is the first research to identify an approach for fake profile identification in LinkedIn. Our approach is based on static profile feature data and not dynamic data, which is not accessible in LinkedIn. *We demonstrate that with limited profile data our approach can identify the fake profile with 84% accuracy and only 2.44% false negative, which is comparable to the results obtained by other existing approaches based on the larger data set and more profile information.*

At present, social network users strongly contemplate on data privacy, in parallel social network communities have advance their security and authentication frameworks to provide better information hiding capabilities to users with new restriction on accessing the information in the network (Fang and LeFevre 2010; Chen and Shi 2009). Along with that this research can be a motivation to work on limited social network information and find solutions to make better decision through authentic data. Additionally, we can attempt similar approaches in other domains to find successful solutions to the problem where the least amount of information is available.

References

Ashcroft, D. and D. Parker (2009). "Development of the Pharmacy Safety Climate Questionnaire: a principal components analysis." Quality and Safety in Health Care 18(1): 28-31.

Bilge, L., et al. (2009). All your contacts are belong to us: automated identity theft attacks on social networks. Proceedings of the 18th international conference on World wide web, ACM.

Boshmaf, Y., et al. (2011). The socialbot network: when bots socialize for fame and money. Proceedings of the 27th Annual Computer Security Applications Conference, ACM.

Bradbury, D. (2011). "Data mining with Linkedin." Computer Fraud & Security 2011(10): 5-8.

Cao, Q., et al. (2012). Aiding the detection of fake accounts in large scale social online services. Proc. of NSDI.

Chakraborty, A., et al. "SPAM: A Framework for Social Profile Abuse Monitoring."

Chen, X. and S. Shi (2009). A literature review of privacy research on social network sites. Multimedia Information Networking and Security, 2009. MINES'09. International Conference on, IEEE.

Cortes, C. and V. Vapnik (1995). "Support-vector networks." Machine learning 20(3): 273-297.

Cristianini, N. and J. Shawe-Taylor (2000). An introduction to support vector machines and other kernel-based learning methods, Cambridge university press.

eBizMBA (2014).

Fang, L. and K. LeFevre (2010). Privacy wizards for social networking sites. Proceedings of the 19th international conference on World wide web, ACM.

Feizy, R., et al. (2009). "Are your friends who they say they are?: data mining online identities." Crossroads 16(2): 19-23.

Fire, M., et al. (2012). "Strangers intrusion detection-detecting spammers and fake profiles in social networks based on topology anomalies." Human Journal 1(1): 26-39.

Günther, F. and S. Fritsch (2010). "neuralnet: Training of neural networks." The R Journal 2(1): 30-38.

Hsieh, C.-J., et al. (2013). Organizational overlap on social networks and its applications. Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee.

Jin, L., et al. (2011). Towards active detection of identity clone attacks on online social networks. Proceedings of the first ACM conference on Data and application security and privacy, ACM.

Joachims, T. (1999). "Making large scale SVM learning practical."

Jolliffe, I. (2005). Principal component analysis, Wiley Online Library.

Kaiser, H. F. (1974). "An index of factorial simplicity." Psychometrika 39(1): 31-36.

Kazienko, P. and K. Musiał (2006). Social capital in online social networks. Knowledge-Based Intelligent Information and Engineering Systems, Springer.

Kontaxis, G., et al. (2011). Detecting social network profile cloning. Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on, IEEE.

Krombholz, K., et al. (2012). "Fake identities in social media: A case study on the sustainability of the Facebook business model." Journal of Service Science Research 4(2): 175-212.

Kumar, A. and D. Zhang (2006). "Personal recognition using hand shape and texture." Image Processing, IEEE Transactions on 15(8): 2454-2461.

Lee, K., et al. (2011). Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. ICWSM.

Li, S., et al. (2004). "Fusing images with different focuses using support vector machines." Neural Networks, IEEE Transactions on 15(6): 1555-1561.

Prieto, V. M., Alvarez, M., & Cacheda, F. (2013). "Detecting Linkedin Spammers and its Spam Nets". International Journal of Advanced Computer Science & Applications, 4(9).

Rheingold, H. (2000). The Virtual Community: Homesteading on the Electronic Frontier, MIT Press.

Riedmiller, M. and H. Braun (1992). RPROP-A fast adaptive learning algorithm. Proc. of ISCIS VII), Universitat, Citeseer.

Shu, W. and Y.-H. Chuang (2011). "Why people share knowledge in virtual communities." Social Behavior and Personality: an international journal 39(5): 671-690.
Stringhini, G., et al. (2010). Detecting spammers on social networks. Proceedings of the 26th Annual Computer Security Applications Conference, ACM.
Taylor, R. (1990). "Interpretation of the correlation coefficient: a basic review." Journal of diagnostic medical sonography 6(1): 35-39.
Xiang, R., et al. (2010). Modeling relationship strength in online social networks. Proceedings of the 19th international conference on World wide web, ACM.
Zhang, C., et al. (2013). "Knowledge-based Support Vector Classification Based on C-SVC." Procedia Computer Science 17: 1083-1090.