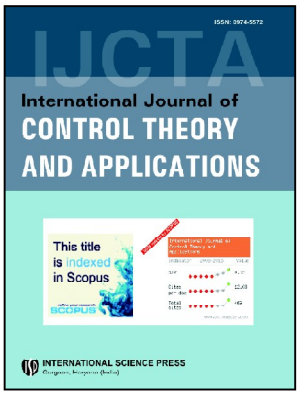

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 29 • 2017

Multistage Fuzzy Classifier Based Phishing Detection Using LDA and CRF Features Followed By Impersonated Entity Discovery

R. Aravindhan[1,*] and R. Shanmugalakshmi[2]

[1] *Department of CSE, Sri Eshwar College of Engineering, Coimbatore, Tamilnadu, India, E-mail: contact2aravind@gmail.com*
[2] *Department of EEE, Govt. College of Engineering, Salem, Tamilnadu, India, E-mail: drshanmi@gct.ac.in*

Abstract: Phishing is aptly defined as an endeavour to grab users' financial and personal information without their knowledge. The information stolen here are their credit card numbers, passwords and social security. All these are achieved through the execution of messaging services and e-mail via electronic communication. The main agenda in proposing this original methodology is for two executions. One is for phishing attacks detection and the other is the recognition of entity/organization that has been exploited by the attackers to execute the phishing attacks. Natural language processing and machine learning – these are the core utilization of the multi-stage methodology proposed. In this methodology the first stage is the discovery of named entities (names of locations, people and organizations) and then the discovery of hidden topics and for this the methods that supports both phishing and non-phishing data i.e. Conditional Random Field (CRF) and Latent Dirichlet Allocation (LDA) is used. Next phase is the AdaBoost stage where the named entities and the hidden topics are treated as features and the messages are classified into phishing or non-phishing. The impersonated entity in the so tracked phishing messages are accomplished through CRF. There arrives no chance for misclassification when <20% is the phishing emails' proportion whilst the phishing attacks is detected by the phishing classifier; as per the perception of the experimental results. 100% - F-measure acquired. Discovery rate is 88.1% in our approach for detecting the impersonated entity from the phishing messages as classified. Any of the legitimate organization may be so mean to the phishing site that is completely offending as the sighting of impersonated entity in phishing is done automatically.

INTRODUCTION

To grab the very confidential information from the targeted individual such as, their credit card details, banking information and passwords the trapper Phishing is used and the targeted individual's good will is attained from showing up us as a legitimate one by impersonating and enticing through the misuse of some other organization's reputation. [1] Financial loss and identity theft are both the results crisis here as the personal information is abused in accessing their account. This sort of phishing lawsuit is initially filed against a Californian teenager (year 2004) and the reason behind the scene way mockery of "America Online" website.

'Phishing' became notorious as the identity stealer and the whole credit goes to nothing other than the advancement of internet in today's era. [2] Personal information theft happens basically by the attackers by

keeping in mind that convincing content is required to trap a user and so the usage of fair email message is being sent; concurrently the hackers troll well throughout the internet. From the sent email's hyperlink, the users will be trapped and tend them to open the page that are totally illegitimate website but in appearance they appear alike the original one. Purchasing a product or existing information updating these are familiar reasons for accusation of financial and personal information. [3-5] A result of such immoral activity ends in deception or selling the valuable information to the seeker in wrong means. It also add one more favour to the attackers and that is they can even add a link implanted in the email and clicking on the same will let the action of downloading malware or malicious code(s). Stealing the trade secrets and accessing the sensitive internal communications of the economic espionage is the very important crime where this sort of crimes becomes very valuable.

Our objective as per proposal is to concentrate both on phishing attack detection and the detection of the organization that is really encouraging the attacker in impersonating them without their knowledge. The main motto of us in developing the robust multi-stage content driven methodology is because it is the one that can detect automatically both phishing messages and the message's impersonated entity; by using it as a filter in the web and email servers. [7] This content filter is built in this methodology by merging the machine learning and natural language processing. For the purpose of feature extraction, the topic discovery methods and named entity extraction is also utilized in this methodology.

- Conditional Random Field (CRF) - Named entities extraction
- Dirichlet Allocation (LDA) - Topic discovery methods

AdaBoost is used for classification and for features topic distribution probabilities and named entities are used and this is how the robust multi stage classifier is developed.

[6]Impersonated entity discovery from the attacks and phishing detection are the areas in which this proposed paper is concentrated on and this shall be termed out as multi-stage methodology.

- *Phishing detection* - identity theft from users is prevented
- *Impersonated organization detection (automatic)* - shutting down the fake website with the knowledge of legitimate organization before those abuses the users.

[8] The companies that in actual fact want their customers to be long lasting should most probably execute this because this is the only approach that keeps their valuable customers safe and secure. This case may be even used little broad minded. That is two or more companies may built a partnership with each other help themselves in eradicating the phishing campaign targeted towards them and this will conclude in benefiting and safeguarding both their customers simultaneously. One of the most remarkable and innovation measure in this proposal/research is the CRF and LDA combined application for detection and discovery. This could be better defined as; the discovery topic discovers the impersonated entity whereas the impersonated entity discovery leads to better topics discovery.

Latest in Phishing Statistics and News

[9-10]As per the IRS report there is an increase in phishing and malware incidents and the percentage is 400 this tax season. This is just a warning from them about the surging of phishing attacks for the consumers.

As per the regard of UK, its Financial Fraud Action has reported many of its findings with reference to phishing. Those are:

- 21% is the increase in percentage per year for the count in deception victims in phishing
- 23% is the increase in percentage per year for the Consumers loss of £174.4 million
- A profile developing in the social media by the Savvy phishers before the victims attack is duly increasing

Information Security professional's survey and mock phishing attack's (millions in count) result are released in the State of the Phish report 2016. Apart from this it's very own key findings are:

- What is the influence of personalization in the open and click rates contributed to end users?
- Industries' mock phishing click rates' average
- Out dated and exceedingly vulnerable plug-ins

75% - this is the increased percentage in one year for US to host 56% of phishing sites as per the strategy of Threat Brief by Webroot. [11] Another stratagem her is the shift come out increased concern in other countries attacks and the case of phishing attacks is targeting US. On top of all as per a report 85,000 IP addresses (malicious) are created in US last year, however, this year is has been terribly increased into 100,000.

A percentage tactics narrated by Vancon Bourne about IT professionals concern about the spear-phishing attacks are:

- 99% of them feel a significant threat will arise against their organization
- 42% among them conclude that it will effect mainly three top-rated organizations
- 28%, this percentage could be the actual phenomenon of spear phishing that their security infrastructure will be influenced from.

Phishing will also focus on the financial institutions and their attack strategies over the same are mentioned in the Easy Solutions report. The little speculative criteria here is that the attacker always wanted a minute i.e. about 190 individuals as targets per attack where they manipulate all their "smash and grab" through malicious sites developed. The thought of downloading the available complete report becomes nullified as registration is mandatory here.

[12] Spoofed pop-up web page this plays a very vital role in trapping the LastPass users and discloses their usernames and passwords. This is some web page that emerges while the LastPass users log into the third-party site. All the above mentioned factors became noticeable by Praesidio (27-year-old CTO) demonstration of the cyber security company. Proof-of-concept has been submitted by him after sending the notice to LastPass that too months before. Spoofing became tougher after this notice as LastPass made a redesigning in their pop-up window. However, phishing hack will find all the service(s) pretty vulnerable that are running within the web browser.

METHODOLOGY

Machine learning and natural language processing are used in this research methodology. [13] And its process is to detect the so happening phishing attacks and the actual reason behind these phishing attacks should be an entity/organization that allows impersonation for the attackers. Multi-stage methodology is segregated into stages.

Stage I

- Named entities (organizations, names of people and locations) extraction
- Hidden topics extraction; phishing and non-phishing data operation through CRF and LDA

Stage II

- Named entities and Hidden topics are considered as features and utilized further
- AdaBoost usage pops out for the classification of all the messages into phishing or non-phishing

Stage III

Usage of CRF is materialized out to determine the impersonated entity in the messages that are concluded or classified as phishing.

Figure 1 illustrates a multi-stage research methodology's schematic representation. This section further narrates the multi-stage research methodology and the AdaBoost, LDA and CRF methods and their reasons for applicability are briefed.

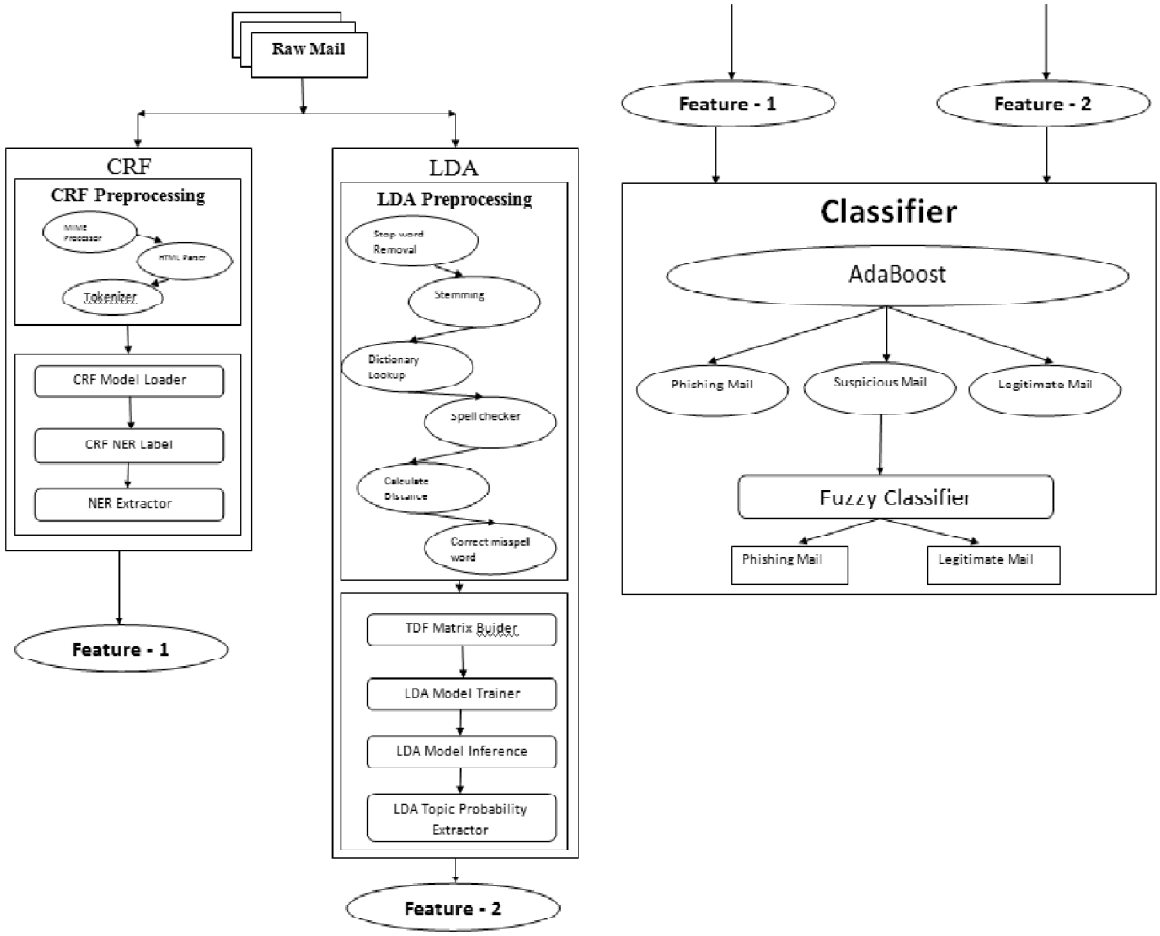

Figure 1: Architecture

Stage I: Feature Extraction (CRF, LDA)

Feature extraction stage is the very first stage. Named entities and topics are the two sets which are the features and they are extracted during this stage. Stage I (a) for CRF used Named entities extraction. Stage I (b) for LDA based topics extraction.

Stage I (a) e named entity feature extraction (CRF)

Phishing detection's classifier is built using the set of features in Stage II. Those set of features are nothing but the named entities extracted in this stage. Names of people, locations, organizations etc are exactly termed out as proper names or proper nouns likely as named entities. From the mailbox of the author the 'phishing email' and 'non phishing email' examples are shown below.

Stage I (b) e feature extraction (LDA)

Feature extraction stage is responsible for extracting the second set of features is comprised as Topics. Formation of topic is the collection of words/phrases. 'Home run', 'Yankees', 'major league' shall be considered as words/phrases for the topic 'baseball'.

The overall compilation of phishing and non-phishing messages leads to the discovery of topics through the help of LDA in the so called Stage 1(b). This topic modelling method LDA has much of the resemblance with the method PLSA in which as per the available documents the hidden theme(s)/topics are identified via bag-of-words approach. Assigning a probability distribution for every document that over all the topics is the privilege of the topic model and it equalises in assigning probability distribution among the words/phrases. PLSA can be defines as LDA's non-Bayesian version. This could be further narrated as, probability theory usage is bit forbidden but the usage of weighted likelihood approach for modelling the latent (hidden) topics are better flourished. Still the usage of PLSA (Blei et al., 2003) leads to two shortcomings and those are: i) Over fitting – a linear growth in PLSA model's parameter number becomes highly possible when the corpus size increases and ii) Apart from training set there occurs requirements for probabilities allocation within the documents but a robust method for allocation is nullified. Dirichlet prior always let us expects its availability in the topic distributions. However, every document owns a detached generative process defined as prescribed by LDA to overcome the above mentioned problems.

Topics that hidden within the phishing messages are discovered by applying LDA is pretty clear here. Illustration of the reason for the LDA usage for phishing is well elaborated through the example (phishing email) taken from CIMB Clicks – an online banking company. 'Financial phishing' topic is comprised from Words/phrases and such Words/phrases are commonly revealed in italic and bolded as well.

Stage II: phishing classifier (AdaBoost)

AdaBoost shown in Fig. 1 is one among the boosting methods and is used in developing a robust classifier that is all about to be shorted about second stage. In the sense of phishing attacks detection, weak and moderately accurate classifiers are put altogether in boosting and thus successfully built a healthy and strong classifier. To bring out the best output in the classification performance, heterogeneous features are combined as well through boosting. Phishing detection's classifier is built strongly through obtaining the topic distribution probabilities and named entities and for that LDA and CRF are used respectively and both of these rely upon Adaboost method. LDA and CRF methods combination is prepared on a particular inspiration and the same is explained here. The documents' words order does not make any reliance over the LDA as it is completely included in the bag-of-words approach. A particular location, organization, name or entity is required always by the CRF for labelling the word(s). Classification stage uses AdaBoost for LDA and CRF fusion. In which CRF is meant for named entity extraction that may help us to overcome the future work and the LDA are given the provisions of prior probabilities for topics discovery technique and that way it could be more organized and improved. When there begins a case for extracting named entity more focussed then the LDA dependent topics discovered shall be used to achieve it.

3.3. Stage III: Fuzzy Classifier

Why Fuzzy Classifier

Soft labelling

Pattern recognition always maintains a criterion theory that the meaning for reciprocally exclusive is none other than the classes. [14] But it slightly differs from the one shown in Figure 1 as an example. Single crisp label (rain) is assigned whole heartedly by the standard classifier. It is absolutely varied from the fuzzy classifier

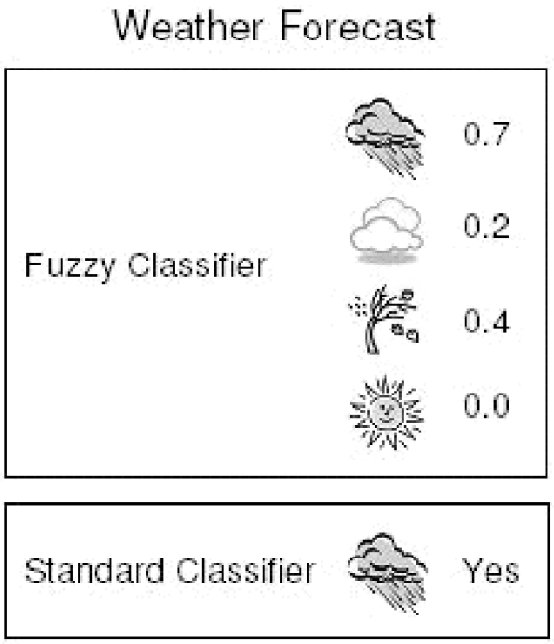

Figure 2: Fuzzy classifiers produce soft class labels

because it assigns soft labels (degrees of membership) that too in all the four classes {rain, clouds, wind, sunshine}. In addition to it, it is accountable for the cloudy weather and winds prospect for the whole day. Soft label offering and posterior probabilities output sounds bit possibility in standard classifier. For instance, consider 0.2 is the possibility of the cloudy weather then we may assume or even declare that there is a chance of 20% for a cloudy day tomorrow. Snow; blizzards or thunderstorms must presume its availability in any one among the active four classes. This valid point is put forth with the assumption of probabilistic model about the formation of full group with four classes. Assumption is a bit lenient while the soft labelling is considered.

$$D: F \rightarrow [0, 1]\, c$$

This is a function approximation as per the observation of soft labels production. Where, F = feature space (place for object descriptions), c = number of classes, D = fuzzy classifier. A spontaneous and useful solution might be provided by the fuzzy classifiers whereas, function approximation regulations to be done remote to the scenario classification becomes quite difficult.

Interpretability

Black box philosophy underpinning classical pattern recognition and political, ethical or legal reasons are main criteria for sidelining the Automatic classification in the applications that are fragmented as challenging for example, medical diagnosis. [15]Intelligible and observable is the status of logic statements and steps that are in the form of proceeding to the class prediction and this is the reason why the designing scenario of the Fuzzy classifiers is considered transparent.

Limited data, available expertise

Terrorist activities, rare diseases, natural disasters, oil depositions are the examples that include the classification and predictions as well. Data, expert opinion or both should be considered while building the Fuzzy classifiers. Keeping the fuzzy classification as the support we are tend to propose a four direct rule generation methods for phishing mail detection. [16] Attributes values' standard deviation and mean is used to generate the fuzzy if-then rules in the first method. Attributes values' histogram is utilized in the second approach for the fuzzy if-then

rules generation. Considering each attribute with a confidence about homogeneous fuzzy sets and this is how the generation of fuzzy if-then rules in the third procedure. [17] Partitions happening between the overlapping areas are what the fourth approach is all about. Training patterns' attribute values information lets the membership function specification of the fuzzy set that is antecedent. This is taken place in the first two approaches where for every class a single fuzzy ifthen rule has been generated. When there is a case occurrence where each attribute consists of homogeneous fuzzy partitions, such fuzzy grids acts as the base for the other two approaches. Every key phishing characteristic indicator's range of values is to be assigned with the high, low, medium linguistic descriptors to commence these methods. Either classes or fuzzy sets classification and consideration are done to the input's valid ranges. As a sample, consider the URL address's length its range obviously varies as „low to „high and even with some in-between values. Specification of apparent boundaries amongst the classes is quite impossible. Degree of membership shall be better described as the selected class and its relevant variable values' degree of belongingness. [18-20] A curve that graphically visualizes the membership value between [0, 1] mapping defined for input space's all point and this is Membership function especially designed for every phishing characteristic indicator. In phishing indicator the linguistic values assigned are low, moderate, and high. When it is for spam mail, then it is Suspicious, Legitimate, Phishy, and Very phishy (trapezoidal membership function and triangular). Input value range is between 0 and 10 and output value range is 0 to 100.

Result

Table 1
LDA Model Performance

Number of topics	*Perplexity*	*Computation time (min)*
5	553.71	1
10	433.36	1
50	260.36	3
100	245.73	6
200	232.27	15

Table 2
Classification (using CRF, LDA, AdaBoost) Performance

% of phishing emails	*True positiverate (TPR)*	*False positiverate (FPR)*	*Precision*	*Recall*	*Fmeasure*	*Area under ROC (AUC)*
50%	0.961	0.039	0.961	0.961	0.961	0.979
40%	0.987	0.013	0.987	0.987	0.987	0.997
30%	0.988	0.012	0.988	0.988	0.988	0.997
20%	1.0	0.0	1.0	1.0	1.0	1.0
10%	1.0	0.0	1.0	1.0	1.0	1.0

Table 3
Impersonated entity discovery (using CRF) performance

Training	*Testing*	*% message with correctly discovered impersonated entity*	*% message with incorrectly discovered impersonated entity*
Data: phishing email Year: 2006 Size: 1 K Source: Phishing Corpus	Data: phishing emails Year: 2006 Size: 1 K Source: Phishing Corpus	86.2	0.0
Data: phishing email Year: 2006 Size: 2 K Source: Phishing Corpus	Data: phishing emails Year: 2012 Size: 1 K Source: SPAM Archive	88.1	0.0
Data: phishing URLs Year: 2010 Size: 25 K Source: Phish Tank	Data: phishing URLs Year: 2010 Size: 25 K Source: PhishTank	74.5	0.0
Data: phishing websites Year: 2011e2012 Size: 15 K Source: crawled	Data: phishing websites Year: 2011e2012 Size: 15 K Source: crawled	81.2	0.0

Tables 1 put forth the results attained through the experiments when multistage architecture is evaluated. Table 1 shows the LDA topics model performance. Temporal strength of LDA is demonstrated through the topics model evaluation on older (2006) and newer (2012). It is well-exposed here that for the 200 topics model, the perplexity is:

- 232.27 – year 2006 dataset
- 873.12 - year 2012 dataset

Phishing classifier is developed using 200 topics model and the reason behind this is there is no reduction in the perplexity regardless to the increased number of topics. LDA's topics distribution probabilities, CRF based named entities extraction and feature set combination led to the fabrication of phishing classifier. Random Forest weak learner is also involved during the development of classifier through AdaBoost. Table 2 presents the 10-fold cross validation results. 50%, 40%, 30%, 20% and 10% is the phishing email's altering proportions in the dataset and the AdaBoost classifier is developed for the same. 50% splits- 0.961, 40% splits - 0.987 and 30% splits - 0.988. This is the classification F-measure acquired for the combined feature set.

Dataset phishing email's varying proportions lead to the variation in the area as 0.979 to 1.0 in the ROC measure (AUC). There are no chances for misclassification and indeed a perfect classification emerges from the AdaBoost classifier when the case is less than or equal to 20% proportion could be maintained the dataset phishing emails. To confirm the classifier's temporal robustness, data from different year were consumed for the obtained results. year 2006 dataset does not consists of the online gaming sites (such as Sulake) and social networking sites (such as Facebook) but to ensure them now we have added newer phishing attacks in the public email corpus (newer) from Jan 2012eFeb 2012 to conduct experiments. Phishing classifier is robust here as the F-measure is 100% due to the dataset's classification performance.

Table 3 shows the impersonated entity discovery model's results. A model that completely comprised of phishing messages is trained using CRF and thus we obtain the above mentioned results. Our approach's outstanding robustness even at data types disparate is well exposed when there are different data types i.e.

phishing URLs, phishing websites and phishing emails used in a model's testing and training. Various years' data are also tested and trained as to our approach's temporal robustness. Training and testing undergone in email datasets has emerged out with the superlative discovery (88.1%) and it is shown in Table 4 as well. It is summarized here as, in the year 2006 emails the model is trained and it is tested in the year 2012 emails and amongst the testing they found about 88.1% of messages that encompass of impersonated organization which is the best discovery happened. Likewise, the strategy of discovery rate as far as our approach gained is phishing URLs - 76.1% and phishing websites - 81.6%. The reason for diminish in the discovery rate when compared with emails is because, the words and the formation of sentences; this is what in which the CRF is dependent upon. Autonomic computing brief overview will be the next section.

CONCLUSION

The main focus and the concept developed in this research is the phishing detection in the robust multi-stage approach and methodology for impersonated entity discovery. Named entities extraction became possible through CRF. Set of features also included the named entities as one among those. LDA provisioned us for the discovery of topics. Another set of features used is the per-document topic probability distributions basically from the LDA topic model. AdaBoost and the combination of named entities and probability estimates are attained the credit of strong classifier construction. The phishing classifier is developed and validated through the consumption of 10-fold cross validation. If lesser then 20% is the phishing emails percentage, then in the test set's boosting method there will be exactly no misclassification in its results. As avoiding the phishing messages towards their users is the first concentrate for all the service providers and as we have developed robust content-driven phishing detection through our investigation they may make use out of this by implemented those in their server side filter. A phishing message's impersonated entity diagnosing will add a magic to the concerned organization as it can even unquestionably warn the entity about the phishing attack which becomes a best solution for all such target attacks. The entity may then deliberately proceed out with the eradicating measures such as, phishing website shall be dragged down and is an innovative measure and other similar operation for keeping their customers safe.

REFERENCES

[1] https://www.ipass.com/press-releases/the-global-public-wi-fi-network-grows-to-50-million-worldwide-wi-fi-hotspots/.

[2] http://www.idc.com/getdoc.jsp?containerId=prUS41061416.

[3] http://www.idc.com/prodserv/smartphone-market-share.jsp.

[4] http://krebsonsecurity.com/2015/08/fbi-1-2b-lost-to-business-email-scams/.

[5] https://blog.cloudmark.com/2016/01/13/survey-spear-phishing-a-top-security-concern-to-enterprises/

[6] Phishing activity trend report, "http://docs.apwg.org/reports/ apwg_trends_report_q1_2013".

[7] Features of phishing e-mail, "http://www.phishing.org".

[8] Phish tank. Available at: http://www.phishtank.com/ accessed on 26 Oct 2009.

[9] B. Kesler, H. Drinan, and N. Fontaine. News briefs. IEEE Security and Privacy, 4(2):8–13, 2006.

[10] S. Garera, N. Provos, M. Chew, and A. D. Rubin. A framework for detection and measurement of phishing attacks. In WORM '07: Proceedings of the 2007 ACM workshop on Recurring malcode, pages 1–8, New York, NY, USA, 2007. ACM.

[11] Y. Pan and X. Ding. Anomaly based web phishing page detection.In ACSAC '06: Proceedings of the 22nd Annual Computer Security-Applications Conference, pages 381–392, Washington, DC, USA, 2006.IEEE Computer Society.

[12] Steve sheng, Brad wardman, Gary warner, Lorrie faith Cranor, Jason Hang, and Chengshan Zhang, "An emprical analysis of phishing blacklists" http://ceas.cc/2009/papers/ceas2009-paper-32.pdf, 2009,Accessed july 2010.

[13] APWG. Anti phishing working group [accessed 30.09.12], http://www.antiphishing.org; 2012.

[14] Asuncion A, Welling M, Smyth P, Teh YW. On smoothing and inference for topic models. In: Proc. of the twenty-fifth annual conference on uncertainty in artificial intelligence, Corvalis, OR; 2009. p. 27e34.

[15] Cortez P, Correia A, Sousa P, Rocha M, Rio M. Spam filtering using network-level properties, advances in data mining, applications and theoretical aspects. LNCS 2010;6171:476e89.

[16] DNSBL. Spam database lookup. Available from: http://www. dnsbl.info/; 2011 [accessed 21.07.11].

[17] Sender ID. Email authentication technology. Available from: http://www.microsoft.com/mscorp/safety/technologies/senderid/default.mspx; 2006 [accessed 21.07.11].

[18] Snort. Network intrusion prevention and detection system. Available from: http://www.snort.org/; 2011 [accessed 21.07.11].

[19] Sperotto A, Vliek G, Sadre R, Pras A. Detecting spam at the network level, the internet of the future. LNCS 2009;5733: 208e16.

[20] SpoofGuard. Available from: http://crypto.stanford.edu/ SpoofGuard/; 2004 [accessed 21.07.11]