Detecting Social Spam Profile on Twitter

Myo Myo Swe, Nyein Nyein Myo
University of Computer Studies, Mandalay (UCSM)
myomyoswe@ucsm.edu.mm, vicky.mdy@gmail.com

Abstract

The fast development of social networking sites such as imparting, sharing, putting away and overseeing huge data leads to pull in cybercriminals. Spammers misuse these social networking sites to abuse cyber laws for their unlawful arts. They start with email, and then quickly spread to new advancements, for example, texting, newsgroups and smart phones. As online social networks, for example, MySpace, Facebook and Twitter turned out to be progressively well known, spammers rapidly found another home for their spamming purposes. Spamming activities of social spammers not only causes dangerous for normal social network users but also annoys to these users. The aim of this paper is to develop social spammer detection approach with low cost and low overhead. The detection approach is a three-phase process: (1) features extraction, (2) features selection and (3) classification. Validation of this approach is tested with 1KS-10KN dataset and CRESCI-2015 dataset.

1. Introduction

People use regularly social networking sites such as Twitter, Facebook, Sina Weibo, etc. and these sites become predominant jobs of their lives. Its advantages are easy to convey, communicate and disseminate information. Twitter allows everyone about their concerns such as their worries, news, jokes, or their inclination. It becomes a viable channel for companies and associations to connect with their clients, advance or offer their items. Because of these benefits, twitter has been progressively utilized for substantial scale data in different fields of human life, for example, promoting news coverage or advertising.

With the well-known of twitter, spammers utilize it for malignant reason; for examples; spreading undesirable troublesome information. Twitter spammers are characterized as noxious users who endeavor to increase social impact and produce spamming substance which contrarily effect on real users. Spammers are also propelled to dispatch different assaults, for example, taking individual data of clients, spreading infections, malware, phishing assaults, or trading off suspicious phony adherents. These issues annoy users, and in addition, cause not only money related misfortune but also security threats. Due to these facts, social spammers are serious problems on Twitter. To become Twitter as a spam free area and enhance the nature of client encounters, we need to portray and recognize social spammers on Twitter.

In this paper, an approach is developed to detect social spammers on Twitter. Profile description of the user is used for detection because most of the spammers are clearly described that they are spammers on their profile descriptions. Random Forest classifier is used for classification because previous researchers used various machine learning classifiers to detect spammers and Random Forest is the best among them. The major contribution of this approach is that a new spam probability feature and twelve account properties features are proposed for spam profile detection. The goal of our detection approach is to detect social spammers with best accuracy and less time complexity. To achieve this goal, the following four steps are needed to accomplish.

- To compute spam probability, this approach uses users' profile description to check whether profile description contains spam or not. Naïve Bayesian classifier is used for calculating the spam probability.
- Account properties features like number of tweets, number of retweet, age of account, number of followers, number of friends, number of favorites, listed count, geo enabled, protected, verified, following rate and follower rate are extracted from the users' profile.
- After features extraction, essential features are chosen by using features selection method. For features selection, Information Gain (IG) is used.

- Random Forest classifier is applied for decision making whether the input profile is spam or not.

The outline of this paper is as follow: In Section 2, background and previous social spammer detection strategies related with our approach are mentioned. In section 3, methodology of the detection approach is briefly explained. Experiment and evaluation are performed in section 4. Section 5 concludes our approach and describes future works.

2. Background and Related Work

Twitter basics functions and existing works in social spammer detection are clarified in this section.

2.1. Twitter

Twitter is an internet based life and advanced news stage that comprises of profiles and a newsfeed. Some basics functions are provided to interact with other users.

- Tweet: The message posted by the users on Twitter is called tweet. 280 characters are allowed to users to post what they want to say.
- Retweet: If a user agrees other user' idea, he or she retweets this tweet to his newsfeed using @RT. Retweet can improve the information dissemination.
- Follow: A person who follows to you is called follower. A person that you follow is called followings. Spammers have many followings and fewer followers.
- Hashtag (#): Users can use hashtag (#) to describe their trending topic.
- Profile Description: Profile description is also called bio. 160 characters are allowed in bio. When one searches another user on Twitter, bio of that user also appears in search result. It likes a showcase and users describe about them that who they are and what they do. Some spammers clearly show that they are spammers. Spammers also embed malicious links such as porn sites, phishing links etc. into profile descriptions. Examples profile description of spammer and normal users are shown in table 1.

2.2. Related Work

In light of the prominence of Twitter, spammers are changing in the rapidly investment on Twitter. To deal with this issue, lots of effective detection strategies are proposed by previous researchers. Benevenuto et. al. [4] examined ascent of video spammers and advertisers in YouTube. Features that could separate spammers from normal users were proposed. Supervised learning strategies were utilized to distinguish video spammers and advertisers. Video-based, user-based and network-based features were extracted for classifying spammers.

Table 1. Examples profile description of spammers and normal users

User	Profile Description
Spammer	2Bbeauty its just 4U Top Israeli beauty cosmetics material from dead sea 4 the most sexy girls worldwide
Spammer	Im a B.Tech Student\r\nI luv 2 explore things & have fun\r\ni luv earning extra money from web :\r\n(PLEASE MAKE @ MENTIONS AFTER YOU FOLLOW ME
Spammer	37 years old\r\nmarried\r\nhave a 12 years old beautiful daugther\r\nVery positive\r\nLots of goals\r\npositive attitude
Spammer	Online Marketing focusing on SEO, Adwords PPC and email marketing.\r\nPlease visit my Web Page: SOLO ADS to 1,450,000 Contacts
Spammer	OK um o a mission to follow at least 1,000 people by 12:00 am and then 10,000 by the end of the week!! startingggggg now!!! ~3B Ent~
Normal	Intentful entrepreneurship, Law of Attraction, The Four Agreements and Personal Growth
Normal	Media Broker & Advertiser at Impacto Digital. Intelligent & Humanist, World Traveler, Speak: Spanish, French & Arabic
Normal	IT Student & Conspiracy Theorist
Normal	I look like a Lindsay, but I roar like a Tigress! Girl Power!
Normal	social media expert, photography hobbyist, i like dealing with people who has a good sense of humor

Amit A. et. Al. [1] implemented CATS system to characterize automation activity of spammers on Twitter. Several novel features that can distinguish spammers from legitimate users were proposed. Fifteen new features like bait-oriented features, behavioral-entropy features, URL-based features, content-based features and profile-based features were extracted for classification. Four machine learning classifiers were applied to characterize spammers. Evaluation was tested on 1KS-10KN dataset. CATS system achieved 93.6% accuracy with Random Forest classifier.

Chkraborty et. al. [2] implemented SPAM system. SPAM system not only monitored but also deleted abusive users from Twitter. It was a framework for social network privacy protection. Twenty features were introduced to detect the abusive activities. The authors trained SPAM on 5000 Twitter users and 200 most recent tweets were used for features extraction. Their approach gave 89% accuracy with support vector machine (SVM) classifier.

Lee et. al. [5] harvested the honeypot accounts to grab spammers. User based features are extracted from the text of tweets posting by the users and these features were added to the classifiers to classify spammers. Validation was performed on 1000 Twitter users. Results of seven classifiers like Decorate, SimpleLogistic, LogitBoost, RandomSubspace, Bagging, J48 and LibSVM were compared. Among these classifiers, Decorate gave the highest accuracy with 88.98%.

User-based features such as number of following, number of followers, reputation, distribution of tweets over 24 hours period and content-based features such as number of URLs, number of replies, keyword weight, number of retweets, average tweet length, number of hashtag are used by McCord et. Al. [6] to detect spammers on Twitter. User-based features are retrieved from user profile and content-based features are retrieved from text of the tweets posted by the users. Random Forest, Support Vector Machine, Naïve Bayesian and K-Nearest Neighbor classifiers were applied for detection. Reputation feature was not useful in their approach because of spammers' evasion. Their approach needed more features like content similarity metric and larger dataset to get the best results.

In ref [7], six user-based features, eleven content-based features and three timing-based features were utilized to detect fake accounts on Twitter. Effective features were chosen by using two feature selection methods: information gain and gain ratio. Random Forest, Decision Tree, AdaBoost, LogistBoost and Bagging classifier were applied. 1KS-10KN dataset was used to train the classifiers. Random Forest was the best classifier and gave 95.7% detection rate.

To detect fake account on Twitter, [8] blacklist corpus was created in advance. For blacklist creation, topic modeling approach and keywords extraction approach were applied. Blacklist based approach achieved better results than the traditional spam words list based approach. In this approach, only user-based features were used and two new features like number of fake words and fake words ratio were proposed. Blacklist based approach utilized Decorate classifier and gave 95.4% accuracy.

Existing works were detected social spammers on Twitter using various approaches. Some detected using blacklist based approach. Some researchers utilize community detection approach. Others use features based approach to detect spammers. Features based approaches such as user-based features, content-based features, automation-based features, URL-based features, behavior-based features and timing-based features were extracted from users' profile and users' tweets. In our approach, profile description is used for spam detection. Because most of the spammers clearly describe about them on profile description. And account properties features and spam probability are combined to determine the input profile is spam or legitimate.

3. Spammer Detection Methodology

In this section, the methodology of spammer detection is briefly described. Our proposed spam profile detection system flow is shown in figure 1. Firstly, in order to classify user into spammer or normal user, standard datasets are needed for classifiers' learning. Firstly, to determine user profile description contains spam words or not, traditional spam classifier (Naïve Bayes) is used for classification. And then, twelve account properties features are extracted. Some features are not useful for detection. Therefore, to achieve effective features, features selection method is used. After effective features are got, classifiers are performed for classifying.

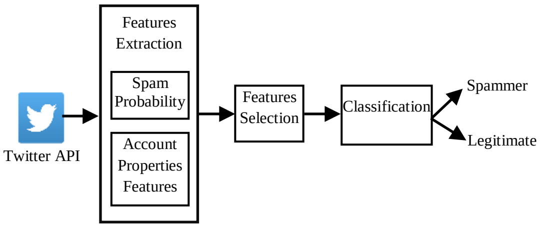

Figure 1. Proposed spam profile detection system flow

3.1. Datasets

In our social spammer detection strategy, two standards datasets: 1KS-10KN dataset and CRESCI-2015 dataset are used.

3.1.1. 1KS-10KN Dataset

Yang et al. [3] crawled the information of 11,000 users and 1354616 tweets from April 2010 to July 2010. 1000 users are labeled as spammers and 10000 users are labeled as normal users. Spammers to normal users ratio is 1:10. Therefore, this dataset is known as 1KS-10KN dataset.

3.1.2. CRESCI-2015 Dataset

CRESCI-2015 dataset is created by Cresci et.al [9]. In this dataset, the data of 3900 users are collected. 1950 users are normal users and 1950 users are spammers. The ratio of spammers to normal users is 1:1, therefore, this dataset is balance dataset.

3.2. Features Extraction

To classify the input profile into spam or legitimate, spam probability feature and twelve account properties features are extracted.

3.2.1. Spam Probability

Profile descriptions of two datasets are used for detection. Profile description includes hashtag(#), mention(@) and links. Hashtag(#), mention(@) and links are removed. Some of words are meaningful and not useful for detection. Stop words (a, and, the, of) are excluded from that data. And then, stemming algorithm is used to find the root or stem of a word. Porter stemmer is utilized for stemming. After stemming algorithm is performed, lemmatization is applied. Lemmatization can give accompanying form; for example; better to good.

The spam detection examines the content of the profile description to detect spam. We have observed that most spam profiles show spam words on their profile descriptions such as free, sex, follow, etc. Most of the spam e-mail filtering systems used Naïve Bayesian because it can correctly predict spam email and take less model building time. Due to these advantages, Naïve Bayesian is used for spam categorization in our approach. To categorize spam, Bayesian is a measurably system which is a powerful strategy. It utilizes earlier information about the preparation tests for step through canny choices about exam tests. This technique computes a likelihood for each profile description utilizing Bayesian insights, as indicated by the tokens in assortment of messages, to verify that a user is spammer or non-spammer. In this way, Bayesian classification computes the corresponding probability based on the content of the profile description P(C = spam|M) to determine profile description belongs to a class, eg., spam. If the probability is over a certain threshold, then the message is from that class. The spam probability of a user can be calculated by the following equation [10].

$$P(spam|M) = \frac{P(M \mid spam)P(spam)}{P(M)} \qquad (1)$$

where,
M = the profile description of a user
C = class (whether spam or real user)

The profile description *M* is represented as a feature vector <f_1, f_2, ..., f_n>, where each feature is one or more words in the profile description [10].

$$P(spam \mid M) = \frac{\left\{ P(spam)\prod_{i=1}^{n} P(f_i \mid spam) \right\}}{\left\{ P(spam)\prod_{i=1}^{n} P(f_i \mid spam) + P(nonspam)\prod_{i=1}^{n} P(f_i \mid nonspam) \right\}} \qquad (2)$$

where,
P(spam|M) = the spam probability of profile description *M*

3.2.2. Account Properties Features

Twelve account properties features are extracted from the user's profile. These twelve account properties features are number of tweets, number of followers, number of friends, number of favorites, listed count, age of account, geo enabled, protected, verified, number of retweet, following rate and follower rate.

(1) Number of tweets: Spammers post more tweets than legitimate users, because they want to persuade legitimate users to click the malicious links. Number of tweets is used as a feature for detection.

(2) Number of followers: Legitimate users post credible information, therefore most of the users follow them and their reputation is high. Legitimate users have more followers than spammers.

(3) Number of friends: Legitimate users build their relationship on trustworthiness. They seldom send friend request to strangers and most of their friends are their family members, their classmates and their colleagues. But, spammers follow to others to get valuable information and want to steal these information. Spammers have higher number of friends.

(4) Number of favorites: The number of favorites of spammers is more than that of legitimate users.

(5) Listed count: Listed count means the number of groups that a user subscribes to. Listed count can segregate spammers and normal users.

(6) Age of account: According to the twitter policy, the age of the legitimate account is at least three months. The more an account is aged, the more it could be viewed as an innocence one.

(7) Geo enabled: Most legitimate accounts are geo enabled or geo localized.

(8) Protected: Most of the legitimate profile are protected.

(9) Verified: Most of the legitimate profile are verified by the Twitter.

(10) Number of retweets: Spammers frequently retweets another users' tweets. They seldom post their original tweets. Spammers have higher number of retweets.

(11) Following rate: This measurement mirrors the speed at which a user pursues different users. Spammers generally have high value of this rate.

$$FollowingRate = \frac{NumberOfFollowings}{Age} \qquad (3)$$

where,
FollowingRate = Following rate
NumberOfFollowings = Number of followings
Age = Age of account

(12) Follower rate: Spammers have low value of follower rate.

$$FollowerRate = \frac{NumberOfFollowers}{Age} \qquad (4)$$

where,
FollowerRate = Follower rate
NumberOfFollowers = Number of followers
Age = Age of account

3.3. Features Selection

There are many features dimensions that are extracted for classification, but, some of the features are not useful for classification. Model building can take long time with many features. To reduce model building time and to choose the effective features, features selection method is used. Information Gain (IG) is used for feature selection. In features extraction phase, spam probability feature and twelve account properties features are extracted from user's profile. Features and their respective weights are shown in table 2. Features that have weight value greater than 0.5 are chosen for detection. Therefore, six features are achieved to detect spam profiles. Spam probability, number of tweets, number of followers, number of favorites, number of friends and listed count are selected for using in classification.

Table 2. Features and their respective weights

Rank	Feature	Weight
1	Spam probability	0.97
2	Number of tweets	0.906
3	Number of followers	0.759
4	Number of favorites	0.665
5	Number of friends	0.564
6	Listed count	0.549
7	Geo enabled	0.478
8	Verified	0.471
9	Protected	0.459
10	Age	0.397
11	Follower rate	0.316
12	Following rate	0.285
13	Number of retweets	0.247

3.4. Classification

Spammer detection is a binary classification task, because it classify a given test Twitter profile into legitimate (+) or spammer (-). Six features are inserted as inputs for classification. To classify spam profile, Random Forest is applied.

Random Forest is a meta-learner which comprises of numerous individual trees. To perform the prediction, the trained random forest algorithm carried out the following steps:

- Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted target or outcome.
- Calculates the votes for each predicted target.
- Considers the high voted predicted target as the final prediction from the random forest algorithm.

The trained random forest algorithm predicts the test data by passing the test features through the rules of each randomly created trees. Each random tree predicts different target (outcome) for the same test feature. Then by considering each predicted target votes will be calculated. Then the final random forest returns the target with highest votes as the predicted class.

4. Experimental Results

The experiment is tested on Core i3 processor, 2GB RAM, 500 GB HDD and 32 bits window 7 OS.

The java programming language (NetBean IDE 8.2) is applied to implement the proposed system. 1KS-10KN dataset, CRESCI-2015 dataset and combined dataset (1KS-10KN+CRESCI-2015) are used for evaluating the proposed system. For testing on 1KS-10KN dataset, 11000 instances are used for evaluation. When testing on CRESCI-2015, 3900 instance are utilized. Evaluating on combined dataset, 14900 instances are used for experiment. We compare the proposed approach with CATS [1] and SPAM [2]. Three metrics such as precision, recall and f-measure are calculated to measure the correctness of our approach. To calculate accuracy, precision, recall and f-measure, this approach uses the following metrics for detection.

True Positive (TP): Legitimate profiles are predicted as legitimate profiles.

True Negative (TN): Spam profiles are classified as spam profiles.

False Positive (FP): Legitimate profiles are classified as spam profiles.

False Negative (FN): Spam profiles are predicted as legitimate profiles.

Accuracy, precision, recall and f-measure are computed as follow.

$$\mathrm{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\mathrm{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$F-measure = 2 \times \frac{\mathrm{Recall} \times \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} \quad (7)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (8)$$

Table 3 shows performance comparison of our approach and two previous approaches (CATS and SPAM). 1KS-10KN dataset is used in CATS and SPAM system. The accuracy of the proposed approach that is tested on 1KS-10KN dataset is 99.96%. When this approach is tested on CRESCI-2015 dataset, the detection rate is 97.18%. The accuracy that are tested on combined dataset is 98.66%. Accuracy of CATS system which is tested on 1KS-10KN dataset is 93.60%. When SPAM system is tested on 1KS-10KN dataset, the accuracy is 89%. Accuracies of the proposed approach that are tested on three datasets are higher than the accuracies of the two previous approaches (SPAM and CATS). The proposed approach achieves best results on three datasets. Accuracy of proposed approach compared with two previous approaches are shown in figure 2. Precision results of three approaches that are tested on 1KS-10KN dataset are shown in figure 3. Figure 4 also compares recall of our technique with that of evaluation done in CATS and SPAM. Compared to the previously proposed approaches, recall of our proposed approach higher than that of two previous approaches. In figure 5, the f-measure of three approaches are compared. F-measure of proposed approach is 1, but f-measure of CATS is 0.932 and f-measure of SPAM is 0.889. Therefore, the proposed approach achieves higher f-measure rate than that of other two approaches.

Table 3. Performance Comparison

	Our Approach (1KS-10KN)	Our Approach (CRESCI-2015)	Our Approach (Combined dataset)	CATS (1KS-10KN)	SPAM (1KS-10KN)
Accuracy	99.96%	97.18%	98.66%	93.6%	89%

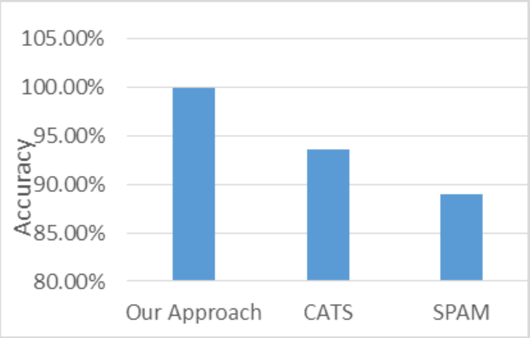

Figure 2. Accuracy comparison

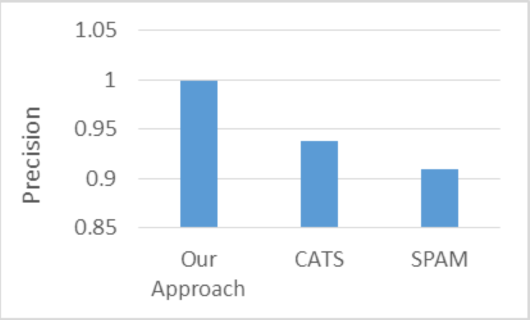

Figure 3. Precision results comparison

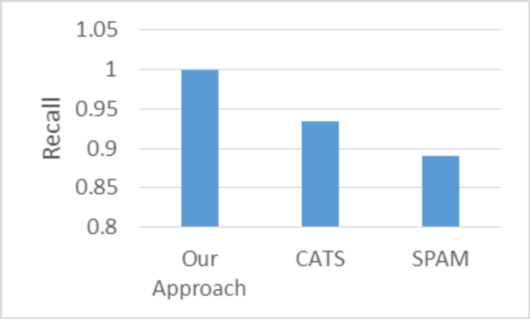

Figure 4. Recall results comparison

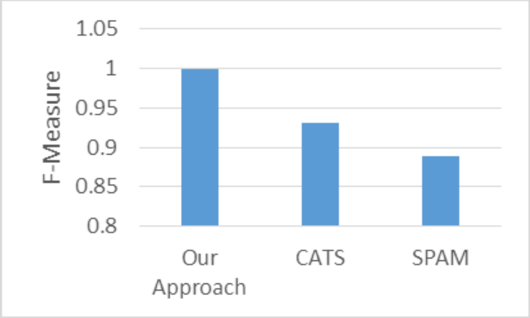

Figure 5. F-Measure results comparison

5. Conclusions

In this paper, we propose a system to detect social spammers on Twitter. The proposed approach use spam probability feature and twelve account properties features to categorize user profile into spam or legitimate. In social spammer detection approach, Random Forest classifier is used for classification. We compare our proposed approach with two approaches (CATS and SPAM). According to these comparison results, it can be seen clearly that the proposed approach achieves better results than the two previous approaches (CATS and SPAM). Implementing features selection method in our approach can not only give best accuracy but also can reduce time overhead. In future, we will discovery new features that are adaptable for other online social networks such as Facebook, Weibo, and etc.

References

[1] Amit A. Amleshwaram, Narasimha Reddy, Sandeep Yadav, Guofei Gu, and Chao Yang, "Cats: Characterizing automation of twitter spammers." *Communication Systems and Networks (COMSNETS), 2013 Fifth International Conference on. IEEE*, 2013.

[2] Ayon Chakraborty, Jyotirmoy Sundi, and Som Satapathy, "SPAM: a framework for social profile abuse monitoring." *CSE508 report, Stony Brook University*, Stony Brook, NY (2012).

[3] Chao Yang, Robert Harkreader, and Guofei Gu, "Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers." *Recent Advances in Intrusion Detection. Springer* Berlin/Heidelberg, 2011.

[4] Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves, "Detecting spammers and content promoters in online video social networks.", *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009.

[5] Kyumin Lee, James Caverlee, and Steve Webb, "Uncovering social spammers: social honeypots+ machine learning." *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 435-442. ACM, 2010.

[6] M. McCord, and M.Chuah, "Spam detection on twitter using traditional classifiers." *International conference on Autonomic and trusted computing. Springer*, Berlin, Heidelberg, 2011.

[7] Myo Myo Swe and Nyein Nyein Myo, "Effective Features Selection for detecting fake accounts on twitter", *The 16th International Conference on Computer Applications (ICCA)*, pp.397-403, February, 2018.

[8] Myo Myo Swe and Nyein Nyein Myo, "Fake Accounts Detection on Twitter Using Blacklist." *The 17th International Conference on Computer and Information Science (ICIS)*, pp. 562-566. IEEE, 2018.

[9] Stefano Crescia, Roberto Di Pietrob, Marinella Petrocchia, Angelo Spognardia and Maurizio Tesconia, "Fame for sale: efficient detection of fake twitter followers", *Decision Support Systems*, 80, pp.56-71.

[10] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?." *IEEE Transactions on Dependable and Secure Computing* 9, no. 6 (2012): pp.811-824.