International Journal of Future Generation Communication and Networking
Vol. 9, No. 1 (2016), pp. 23-34
http://dx.doi.org/10.14257/ijfgcn.2016.9.1.03

Cloud Based Forensics Framework for Social Networks and A Case Study on Reasoning Links between Nodes

Fawang Han

School of information technology, Nanjing Forest Police College, Nanjing, Jiangsu, China, 210046
hanfawang@163.com

Abstract

Recently, social networks have become one of the most popular tools for communication and information exchanges, and people are constructing social relationships and conducting social interactions over social networks. In this paper, we focus on digital forensics on social networks. Specifically, considering the emerging of cloud computing and big data tides, we propose a cloud based forensics framework for social networks, where social networking data is collected, stored, and analyzed through a multi-layered modularity framework using cloud computing techniques, including virtualization, distributed processing and storage, and collaboration. Besides, we also provide a case study on social link prediction, which is, reasoning links between nodes to infer possible relationships between criminals.

Keywords*: Digital forensics, Social networks, Cloud computing, Link prediction*

1. Introduction

With the rapid development of computer and Internet technology, committing crimes through digital channels becomes even more imperceptible. However, more and more criminal evidences would be included, and therefore digital forensics technique takes a significant role in computer based crime detection and control [1].

Generally, digital forensics is a process of acquiring, storing, analyzing and archiving digital evidences. As an important part of digital forensics, network forensics uses network technology to deal with network crimes, which refers to crimes committed through network such as attacks or intrusions over network systems or information. Network crimes are typically intelligent, concealed, complicated and anonymous, and could cause property loses and even endanger public safety and national security. Therefore, the significance of network forensics is perceived. Typical process of network forensics is shown as Figure 1 [2].

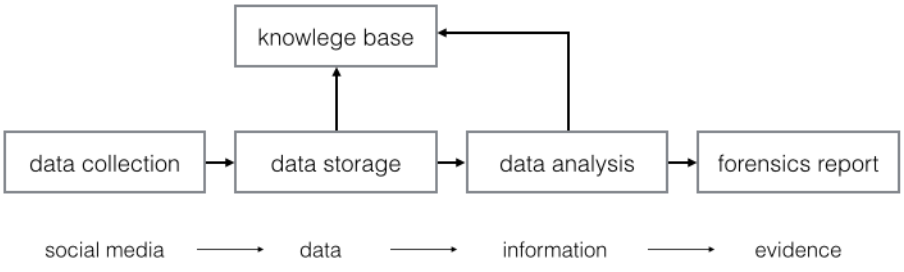

Figure1. Typical Process of Network Forensics

ISSN: 2233-7857 IJFGCN

Recently, social networks have become one of the most popular tools for communication and information exchanges, and people are constructing social relationships and conducting social interactions over social networks. Therefore, in this paper, we focus on forensics on social networks. That is, we collect information from social networking sites, analyze social networking data and try to infer some useful evidence for crime control.

Moreover, the emerging of cloud computing [3] and big data [4] tides has made the existing forensics methods difficult due to the dynamics and scale of massive logs and datasets. Fortunately, the characteristics of cloud computing such as open standard, connected collaboration, and rapid and secure storage and computing services, also facilitate the forensics process if customized forensics model based on cloud computing is utilized.

To this end, in this work, we propose a cloud based forensics framework for social networks, where social networking data is collected, stored, and analyzed through a multi-layered modularity framework using cloud computing techniques, including virtualization, distributed processing and storage, and collaboration. Besides, we also provide a case study on link prediction [5]. That is, reasoning links between nodes to infer possible relationships between criminals.

The reminder of this paper is organized as follows. Section 2 presents related work. In Section 3, we discuss the proposed cloud based forensics framework for social networks in details. Then, Section 4 provides a case study on link prediction based on proposed forensics framework. Experiments are conducted in Section 5, and finally Section 6 concludes the paper.

2. Related Work

The first category of related work is web based network forensics, which aims to obtain the web browsing data for analysis. Typically, there are three methods. The first method is server end web forensics. Wu [6] designed a dynamic forensics method for ASP websites. However, the forensics becomes even more difficult and expensive ever since the growing of cloud computing clusters [7]. The second method is client end forensics. The major issue is to analyze the logs of all the possible related software in details [8]. The last method is data stream based forensics. For example, Shanmugasundaram [9] presented a distributed forensics network called ForNet. However, this kind of forensics is typically difficult to implement. In this work, we employ a client end forensics method by actively crawling data using a cloud based forensics framework.

The second category of related work is social network forensics. Son [10] investigated evidence extraction tools to measure the capability of extracting evidence from SNSs in different test scenarios, and identified current issues and limitations of the tools. Mulazzani [11] discussed the important data sources and analytical methods for the forensic analysis of social networks, and then demonstrated using a Facebook case study. Cheng [12] developed tools for installation on a user's computer to provide them the ability to retrieve other online user information via chat and social network websites. Markus [13] collected social networking data based on a custom add-on for social networks in combination with a web crawling component. In this work, inspired by [13], we integrate the crawling component with a cloud based infrastructure for social network forensics.

The last category of related work is link prediction in social networks. Indeed, link prediction is a well-studied problem [5]. Bao [14] used principal component analysis to identify features that are important to link prediction. Li [15] proposed a link prediction method based on clustering and global information. Xie [16] proposed to extract user interest topic feature and network topology structure feature, and then fed into a SVM

classifier to predict possible links. In this work, after presenting the cloud based forensics framework, we provide link prediction on social networks as a case study, given the assumption that links between criminals could be useful evidences for crimes control.

3. Cloud Based Forensics Framework for Social Networks

As mentioned earlier, we focus on the network forensics problem in the field of social networks, and embrace cloud computing techniques. In this section, we present our cloud based forensics framework for social networks.

Figure 2 gives the illustration of our proposed forensics framework. Generally, the underlying infrastructure is built upon cloud computing suites such as Hadoop [17], and virtualization techniques are leveraged for multiple user operation and data storage. Then, a crawler is constructed for collecting social network data, which is afterwards fed into the data analysis component that outputs potential useful evidences.

As shown in Figure 2, we have five layers in the forensics framework: infrastructure layer, virtualization layer, data pool layer, crawler layer and analysis layer. Now we describe the components and functions of each layer in details.

(1) Infrastructure layer: includes the underlying infrastructure, such as data nodes, storage and network facilities. Specifically, we employ Hadoop as our infrastructure base, which provides storage, computing and network services for upper layers.

(2) Virtualization layer: includes multi-tenant structure, which allows for data separation and sharing; parallel and distributed process, which provides multi-thread services, distributed cache, and large scale capability; and log management, which standardizes logs achieved by static or dynamic forensics methods, and prepares logs for further analysis.

(3) Data pool layer: stores data separately, including user log files, system log files, network log files, attack log files, and update log files. Note that data in this layer is used for management instead of crawled web pages.

(4) Crawler layer: is the most important layer in the framework. Unlike above three layers, which focus on the infrastructure and management perspective, this layer is responsible for social network data collection. Crawler layer includes three main components: user authentication and access control, typically related to the rules of specific social network sites; task and resource scheduling and management, which controls the workflow of crawler structure; and download, parse and store crawled data, which deals with the web pages directly and might be involved with the open API of specific social network sites.

(5) Analysis layer: is application oriented and includes log query, management and mining, which analyzes logs and provides insights in terms of logs; social network analysis, which analyzes social network data crawled from lower layers; and other forensics application based on logs and social network data. Specifically, we introduce Hadoop Mahout for data mining and analysis to mine potential evidences.

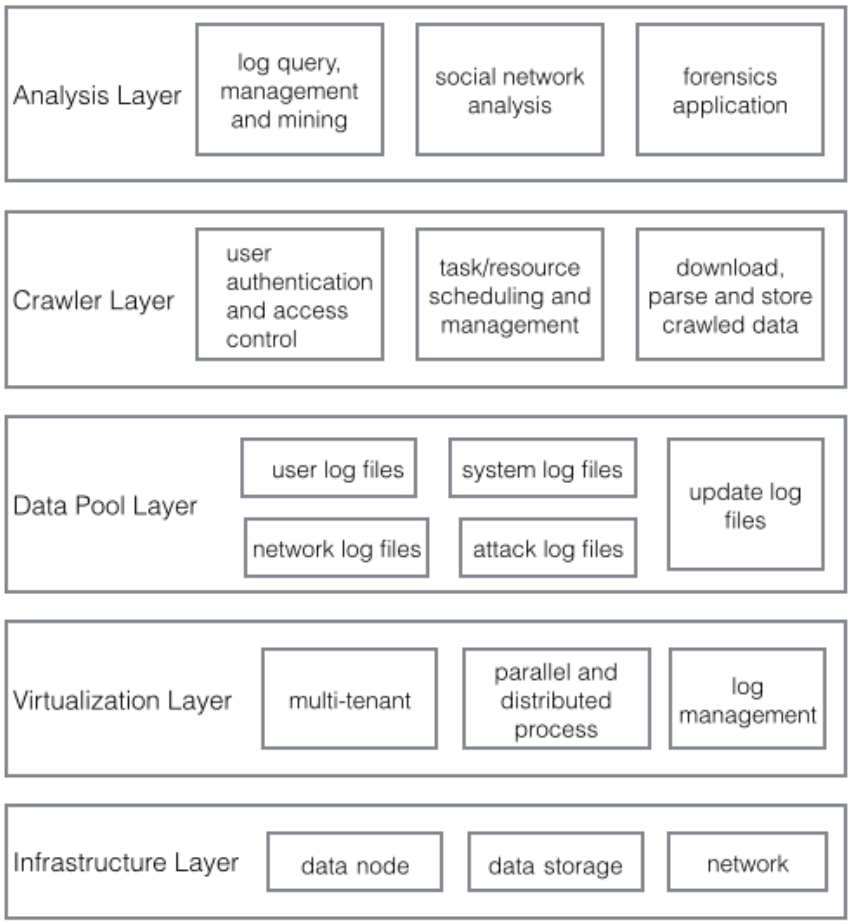

Figure2. Multi-Layer Cloud Based Forensics Framework for Social Networks

Moreover, Figure 3 gives the structure of distributed crawler. First, multiple crawlers are managed by a controller node, which is responsible for starting, stopping and scheduling crawlers. Then, collected data of each crawler is transferred into HDFS storage. HDFS controller node assigns data blocks onto different DataNodes and JobTracker node schedules different TaskNodes for job execution. Note that we simplify the deployment by putting each DataNode and TaskNode onto the same physical node. The job scripts are defined upon business applications, and the data results are stored into a business database for further use.

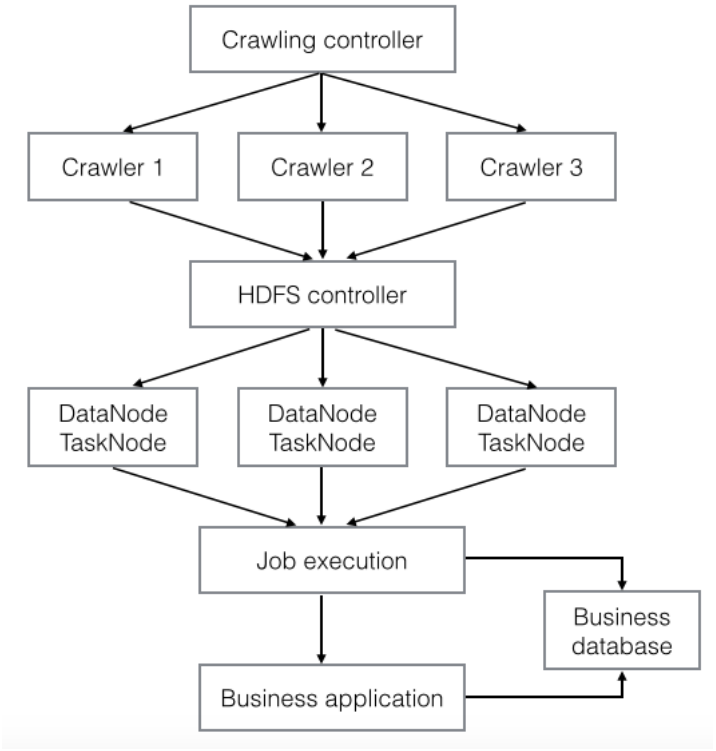

Figure3. Structure of Distributed Crawler

4. Case Study: Reasoning Links between Nodes

In this section, we discuss a typical case study on social network data forensics using proposed cloud based forensics framework. Indeed, if all the social links between criminals are obtained, we could infer potential relationships between criminals or even predict who else could be a potential criminal. Therefore, we believe that reasoning social links between nodes is one of the most significant forensics applications in social networks.

Figure 4 describes the overall model for reasoning links between nodes. Basically, the reasoning model is built upon three categories of features: a) network structure similarity, b) user interests similarity and c) check-in places similarity. The overall workflow is as follows. First, construct the social network between users, and then extract user interests information from nodes. After that, based on the extracted check-in information from social network sites, calculate the location based similarity between users. By integrating above three kinds of features into a classifier model, the existence of social links can be predicted given specific node.

Now we discuss each category of features one by one. First, capturing structural characteristics of social networks can improve the accuracy of link prediction. One intuition is that more common friends two nodes have, more similar they are. Therefore, the structural similarity is positive related to the number of common friends. Suppose c is a common friend of nodes a,b. Another observation is that if c is linked to many other nodes, the strength of connection between a,b transitioned though c is relatively weak. That is, the larger degree of c has, the less strength of the social link between a,b is. Based on above two observations, we define the structural similarity between a,b as:

$$struc\_score(a,b) = \sum_{c \in N(a) \cap N(b)} \frac{1}{\log \deg(c)}, \quad (1)$$

where $N(a), N(b)$ denote the neighbors of a, b respectively, and $\deg(c)$ is the degree of node c. Apparently, we can see that the summation denotes the social link is positively related to the number of common neighbors, and the fraction denotes it is negatively related to the degree of common neighbors.

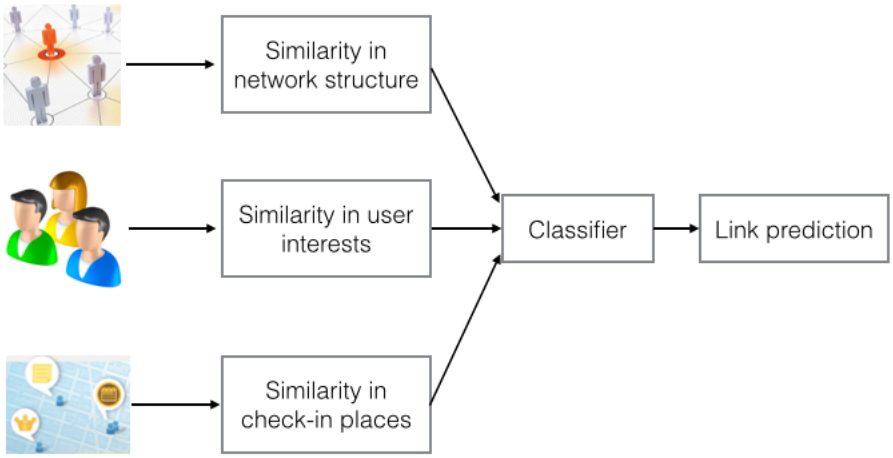

Figure 4. Model for Reasoning Links between Nodes

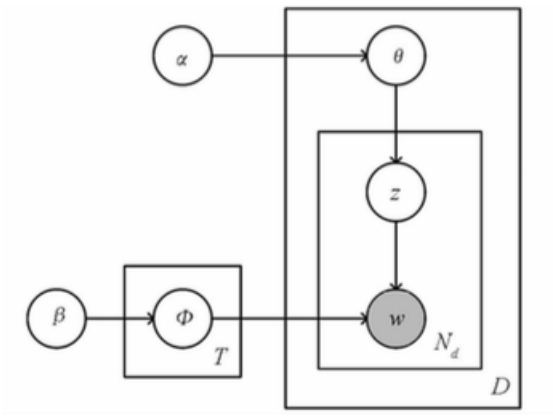

Figure 5. Graphical Representation of LDA

Second, other than structural characteristics, semantic information or topic-based characteristics are also significant in social networks. We consider leveraging Latent Dirichlet Allocation (LDA) topic model [18] for learning potential semantic information from text.

Basically, LDA is a common used topic model, which can learn latent topics from a corpus of documents. It views each document as a probability distribution over a set of topics, and each topic as a probability distribution over a set of words. Figure 5 shows the graphical representation of LDA model, and Table 1 lists the notations. Each document is associated with a set of topics with a multinomial distribution q, and each topic is associated with a set of words with a multinomial distribution F. For the generation of each word w in document d, first extract a topic z from q, and then extract a word w from F. Repeat the process N_d times, where N_d is the number of words in d, we get document d.

Table 1. Notations in LDA

Notation	Description
d	A document
w	A word
z	A topic
D	The set of documents
W	The set of words
T	The set of topics
q	Multinomial distribution of document-topic
F	Multinomial distribution of topic-word
a	Multinomial parameter for q
b	Multinomial parameter for F

User interests similarity is defined upon the latent topics learned from users. Specifically, first learn the latent topics from the textual posts of users, and secondly, construct the topic vectors and calculate the similarity between topics vectors as the user interests similarity feature. Taking Twitter as an example, to solve the short text issue, we take the whole history of user posts as one single document instead of each single tweet, and the set of all users as the corpus.

Let corpus be $D=\{d_1,d_2,...,d_N\}$, where N is the number of documents (i.e., users), and each document be $d_i=\{w_{i1},w_{i2},...,w_{iC_i}\}$, where C_i is the number of words in d_i. After learning using LDA, we get the document-topic distribution $q=\{p_{ij}\}$ and topic-word distribution $F=\{q_{ij}\}$. Suppose the topic vectors for users d_a,d_b are $q_a=\{p_{a1},p_{a2},...,p_{aT}\}$ and $q_b=\{p_{b1},p_{b2},...,p_{bT}\}$ respectively, where T is the number of topics. Therefore, the similarity between them is calculated as:

$$topic\_score(a,b) = \cos(q_a, q_b) = \frac{\sum_{j=1}^{T} p_{aj} \cdot p_{bj}}{\sqrt{\sum_{j=1}^{T} p_{aj}^2 \sum_{j=1}^{T} p_{bj}^2}}. \quad (2)$$

The larger $topic\_score(a,b)$ is, the more similar user interests are. If the topic interests between users are similar, it is more likely that there exists a link between them.

The last feature is check-in places similarity. Since more and more social networking sites are Location-Based Services (LBS), which associate the time sequences and behavior routes with geographical locations to connect social network users with real world, check-in places as one of the most significant features of LBS are also important for link prediction. Indeed, check-in places can reflect user preference in a geographical way. If the check-in places between two users are similar, it is more likely that there exist a link between them.

Suppose the check-in sequence for user a is $(l_{a1}, l_{a2}, ..., l_{aN_a})$, where N_a is the number of check-in places for a, and the check-in times for all locations are $(c_{a1}, c_{a2}, ..., c_{aN_a})$; and the check-in sequence for user b is $(l_{b1}, l_{b2}, ..., l_{bN_b})$, where N_b is the number of check-in places for b, and the check-in times for all locations are $(c_{b1}, c_{b2}, ..., c_{bN_b})$. Let $Dist(l_{ai}, l_{bj})$ be the distance between locations l_{ai}, l_{bj}. In this work, we simply consider locations within 0.5 km as the same places. And therefore, the check-in places similarity is defined as:

$$loc\_score(a,b) = \sum_{i=1}^{N_a} \sum_{\substack{j=1 \\ if\ Dist(l_{ai}, l_{bj}) < 0.5}}^{N_b} (c_{ai} + c_{bj}) \Big/ \left(\sum_{i=1}^{N_a} c_{ai} + \sum_{j=1}^{N_b} c_{bj}\right). \quad (3)$$

Now we have three features as Equations (1) (2) and (3), representing user similarities in network structure, semantic topic and check-in places respectively. Next we integrate all three features into a classifier for link prediction. In this work, we use Support Vector Machines (SVM) and Bayesian statistics model respectively as the classifier. Specifically, we label the existing links among user nodes as positive instances, and user pairs with unknown links as testing data.

5. Experiment

In this section, we evaluate the link prediction between user nodes. The dataset is collected through our proposed forensics framework and social networking crawler. In our experiment, the data is collected from Twitter, one of the most popular social networking sites, which provides an open API and facilitates the data collection process. We implement SVM and Bayes model using LibSVM [19] and PyMix [20] respectively.

First, we validate the efficiency of extracted three features. Specifically, we use Bayes model reproduce the topology of social network. That is, given three features, notated as sc, tc, lc for $struc\_score, topic\_score, loc\_score$, we calculate the probability of users being friends. Suppose given feature sc, the probability of existing link between users a, b can be calculated as:

$$P(ab \mid sc) = \frac{P(sc \mid ab)P(ab)}{P(sc)}, \qquad (4)$$

where $P(ab)$ is the ratio of the number of friend links to the total number of edges in the network, $P(sc \mid ab)$ is the probability of friend links with feature sc, and $P(sc)$ is simplified as the probability of non-friend links with feature sc.

Suppose three features are independent, we have

$$P(ab \mid sc,tc,lc) = P(ab \mid sc) + P(ab \mid tc) + P(ab \mid lc). \qquad (5)$$

Figure 6 gives the ROC curve of extracted features using Bayes model. We can observe that the area under ROC curve for integrating three features is 0.8523, which indicates that the extracted three features can represent the dataset well. Besides, we can see that location feature only performs worst, which means that location feature should be used with other kinds of features for best performance.

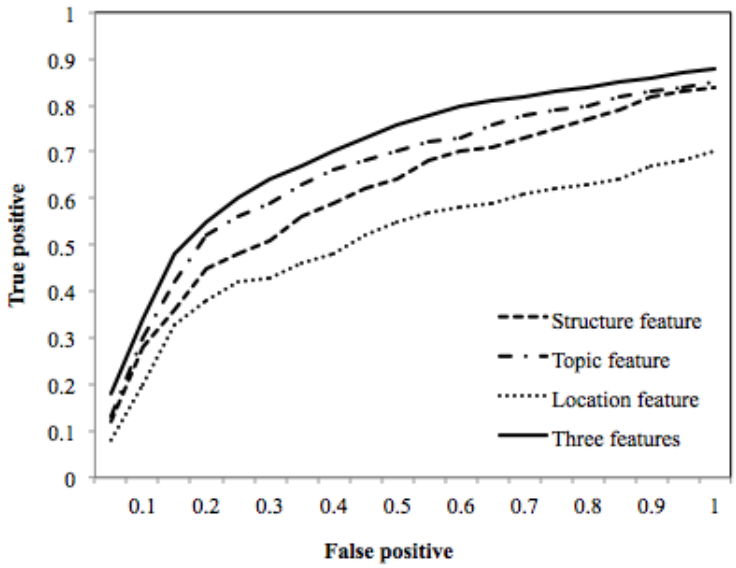

Figure 6. ROC Curve of Extracted Features

Table 2. Results of Structural Link Prediction

Model	Training/test set	Precision	Recall	F-measure
SVM	8:2	0.9028	0.6512	0.7566
	7:3	0.8845	0.6127	0.7239
Bayes	8:2	0.8979	0.6486	0.7532
	7:3	0.8697	0.5989	0.7093

Table 3. Results of Temporal Link Prediction

Model	Training/test set	Precision	Recall	F-measure
SVM	8:2	0.8895	0.6334	0.7399
	7:3	0.8243	0.5892	0.6872
Bayes	8:2	0.8769	0.6211	0.7272
	7:3	0.8126	0.5778	0.6754

Moreover, we evaluate the performance of link prediction by two sets of experiments. The first one is predicting links from the perspective of network structure, and the second experiment is predicting temporal links with consideration of timestamps of link formation. That is, structural link prediction only evaluates the existence of potential links, while temporal link prediction also considers the time sequence of social links, which means to predict future links based on history links. In order to evaluate the accuracy of link prediction, we generate test set by randomly deleting existing links from data collection.

Tables 2 and 3 show the results of structural and temporal link prediction using SVM and Bayes model respectively. We evaluate the prediction results using precision, recall and F-measure with the ratio of training data and test data as 8:2 and 7:3. We have the following observations. First, although temporal link prediction is slightly worse than structural prediction, our method can also produce satisfactory results for reasoning potential future links. Second, SVM based link prediction model is slightly better than Bayes based model. Third, the more training data we use, the more accuracy the prediction is.

6. Conclusion

In this paper, we provide a preliminary effort on forensics research for social networks. Specifically, we propose a forensics framework based on cloud computing infrastructure and web crawling component. Then, using collected data, we provide a case study on link prediction. In future works, we would like to investigate more applications on forensics analysis using social networking data.

Acknowledgment

This paper is supported by the year in 2015, Nanjing Forest Police College, “The Fundamental Research Funds for the Central Universities”. The project number is LGYB201505

References

[1] F.N. Dezfoli, A. Dehghantanha and R. Mahmoud, “Digital Forensic Trends and Future [J]”, International Journal of Cyber Security & Digital Forensics, **(2013)**.
[2] E. Casey, “Handbook of digital forensics and investigation”, Handbook of Digital Forensics & Investigation, **(2011)**.
[3] Q. Zhang Q, L. Cheng and R. Boutaba, “Cloud computing: state-of-the-art and research challenges[J]”, Journal of Internet Services & Applications, vo. 1, no. 1, **(2010)**, pp. 7-18.
[4] S. Sagiroglu and D. Sinanc, “Big data: A review”, //International Conference on Collaboration Technologies & Systems. IEEE, **(2013),** pp. 42 - 47.

[5] M.A. Hasan and M.J. Zaki, "Social Network Data Analytics", Springer US, **(2011),** pp. 243-275.
[6] C.S.Wu and M. Qiu, "The method for obtaining electronic evidence from ASP dynamic website", Forensic Science and Technology, vol. 5, **(2010)**, pp. 43-45.
[7] K.Y. Ruan, J. Carthy, T. Kechadi and M. Crosbie, "Cloud forensics: An overview", In: Proc. of the Advances in Digital Forensics VII, **(2012)**, pp. 15-25.
[8] L.D. Zhong, "Forensic analysis of Web browser with dual layout engine", In: Proc. of the 1st Int'l Conf. on Digital Forensics and Investigation (ICDFI), Beijing, **(2012)**.
[9] K. Shanmugasundaram, N. Memon and A. Savant, "ForNet: A Distributed Forensics Network", Lecture Notes in Computer Science, **(2003),** pp. 1-16.
[10] J. Son, D. Forensics and S.N. Forensics, "Social Network Forensics", Lap Lambert Academic Publishing, **(2012)**.
[11] M. Mulazzani, M. Huber and E. Weippl, "Advances in Digital Forensics VIII", Springer Berlin Heidelberg, **(2012)**.
[12] J. Cheng, J. Hoffman and T. Lamarche, "Forensics Tools for Social Network Security Solutions", Pace Univ, **(2009)**.
[13] M. Huber, M. Mulazzani, M. Leithner, S. Schrittwieser, G. Wondracek and E. Weippl, "Social snapshots: digital forensics for online social networks. In*Proceedings of the 27th Annual Computer Security Applications Conference* (ACSAC '11)", ACM, New York, NY, USA, **(2011)**, pp. 13-122, DOI=10.1145/2076732.2076748.
[14] Z. Bao, Y. Zeng and Y.C. Tay, "sonLP: social network link prediction by principal component regression [J]", IEEE/ACM International Conference on Advances in Social Networks Analysis & Mining, **(2013)**, pp. 364 - 371.
[15] F. Li, J. He and G. Huang, "A Clustering-based Link Prediction Method in Social Networks", Procedia Computer Science, **(2014)**, pp. 432-442.
[16] X. Xie, Y. Li and Z. Zhang, "A Joint Link Prediction Method for Social Network", Communications in Computer & Information Science, **(2015)**.
[17] C. Lam, "*Hadoop in Action* (1st ed.)", Manning Publications Co., Greenwich, CT, USA, **(2010)**.
[18] D.M. Blei, A.Y. Ng and M.I. Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research, **(2003)**.
[19] C. Chang and C. Lin, "LIBSVM: a Library for Support Vector Machines", Acm Transactions on Intelligent Systems & Technology, vol. 2, no. 3, **(2001)**, pp. 389-396.
[20] B. Georgi, I.G. Costa and A. Schliep, "PyMix - The Python mixture package - a tool for clustering of heterogeneous biological data", Bmc Bioinformatics, vol. 11, no. 2, **(2010)**, p. 9.

