ERCIM "Alain Bensoussan"
Fellowship Scientific Report

Fellow: Wei Wang
Visited Location: SnT Centre, University of Luxembourg
Duration of Visit: 12 months, 14/01/2010 – 13/01/2011 (two exchange programs included)

I - Scientific activity

During the period of the fellowship, I focused on three research directions: anomaly intrusion detection, network traffic measurement and recommender systems.

In the first topic, I built lightweight intrusion detection models by extracting *exemplars* from the training data and by selecting important *attributes* from the data, so that the original data can be represented by only a smaller set of data samples or of the most important attributes while redundant samples and unrelated attributes are discarded. This procedure can be regarded as data abstraction. In our work, we used new clustering algorithm like Affinity Propagation (AP) and traditional one like *k*-means to extracted *exemplars* from the training data. We used Principal Component Analysis (PCA) for abstracting *attributes* from the data. We also used Information Gain to select the important *attributes* from the data. Extensive experiments were conducted based on real HTTP traffic as well as on KDD'1999 data and the test results showed that data abstraction is able to significantly improve the detection efficiency as well as detection accuracy. Two related papers have been accepted.

I have worked on the topic of network traffic analysis during my visit to NTNU. As a joint work with Q2S/NTNU, we developed a Robust PCA-based traffic anomaly detection method using Netflow traffic as the data source collected from UNINETT. I am very interested in this topic. I have co-organized a Special Issue on "Network Traffic Monitoring and Analysis" for IEEE Network. I am currently working on this topic and more results are expected in the future. A related paper has been published.

For the third topic, I have collaborated with Dr. Georgios Pitsilis who is also an ERCIM fellow at the University of Luxembourg. We explored the benefits of combining clustering and social trust information for Recommender Systems. The test results showed that the potential advantages in using clustering can be enlarged by making use of the information that social networks can provide. A related paper has recently been accepted by IFIP Trust Management 2011.

II- Publication(s) during your fellowship

1. Wei Wang, Xiangliang Zhang, "High-speed Web Attack Detection through Extracting Exemplars from HTTP Traffic", to appear in the proceedings of 26th ACM Symposium on Applied Computing (**SAC' 2011**) (Security Track), TaiChung, Taiwan, ACM Press, 21-25 March, 2011.

Abstract: In this work, we propose an effective method for high-speed web attack detection by extracting exemplars from HTTP traffic before the detection model is built. The smaller set of exemplars keeps valuable information of the original traffic while it significantly reduces the size of the traffic so that the detection remains effective and improves the detection efficiency. The Affinity Propagation (AP) is employed to extract the exemplars from the HTTP traffic. K-Nearest Neighbour (k-NN) and one class Support Vector Machine (SVM) are used for anomaly detection. To facilitate comparison, we also employ information gain to select key attributes (a.k.a. features) from the HTTP traffic for web attack detection. Two large real HTTP traffic are used to validate our methods. The extensive test results show that the AP based exemplar extraction significantly improves the real-time performance of the detection compared to using all the HTTP traffic and achieves a more robust detection performance than information gain based attribute selection for web attack detection.

2. Xiangliang Zhang, Wei Wang, Kjetil Nørvåg, Michele Sebag, "K-AP: Generating Specified K Clusters by Efficient Affinity Propagation", 10th IEEE International Conference on Data Mining (**ICDM' 2010**), pp. 1187-1192, Sydney, Australia, December 14-17, 2010 (acceptance rate=155/797=19.4%).

   Abstract: The Affinity Propagation (AP) clustering algorithm proposed by Frey and Dueck (2007) provides an understandable, nearly optimal summary of a data set. However, it suffers two major shortcomings: i) the number of clusters is vague with the user-defined parameter called self-confidence, and ii) the quadratic computational complexity. When aiming at a given number of clusters due to prior knowledge, AP has to be launched many times until an appropriate setting of self-confidence is found. The re-launched AP increases the computational cost by one order of magnitude. In this paper, we propose an algorithm, called K-AP, to exploit the immediate results of K clusters by introducing a constraint in the process of message passing. Through theoretical analysis and experimental validation, K-AP was shown to be able to directly generate K clusters as user defined, with a negligible increase of computational cost compared to AP. In the meanwhile, KAP preserves the clustering quality as AP in terms of the distortion. K-AP is more effective than k-medoids w.r.t. the distortion minimization and higher clustering purity.

3. Wei Wang, Xiangliang Zhang, Georgios Pitsilis, "Abstracting Audit Data for Lightweight Intrusion Detection", 6th International Conference on Information Systems Security (**ICISS' 2010**), pp. 201-215, Springer, Gandhinagar Gujarat, India, 15-19 December 2010 (acceptance rate =14/51=27.5%).

   Abstract: High speed of processing massive audit data is crucial for an anomaly Intrusion Detection System (IDS) to achieve real-time performance during the detection. Abstracting audit data is a potential solution to improve the efficiency of data processing. In this work, we propose two strategies of data abstraction in order to build a lightweight detection model. The first strategy is exemplar extraction and the second is attribute abstraction. Two clustering algorithms, Affinity Propagation (AP) as well as traditional k-means, are employed to extract the exemplars, and Principal Component Analysis (PCA) is employed to abstract important attributes (a.k.a. features) from the audit data. Real HTTP traffic data collected in our institute as well as KDD 1999 data are used to validate the two strategies of data abstraction. The extensive test results show that the process of exemplar extraction significantly improves the detection efficiency and has a better detection performance than PCA in data abstraction.

4. Atef Abdelkefi, Yuming Jiang, Wei Wang, Arne Aslebo, Olav Kvittem, "Robust Traffic Anomaly Detection with Principal Component Pursuit", **ACM CoNEXT' 10 Student Workshop**, Philadelphia, USA, November 2010.

   Abstract (Introduction part): Principal component analysis (PCA) is a statistical technique that has been used for data analysis and dimensionality reduction. It was introduced as a network traffic anomaly detection technique in 2004. Since then, a lot of research attention has been received, which results in an extensive analysis and several extensions. In literature, the sensitivity of PCA to its tuning parameters, such as the dimension of the low-rank subspace and the detection threshold, on traffic anomaly detection was indicated. However, no explanation on the underlying reasons of the problem was given. Further investigation on the PCA sensitivity was conducted in related work and it was found that the PCA sensitivity comes from the inability of PCA to detect temporal correlations. Based on this finding, an extension of PCA to Kalman-Loeve expansion (KLE) was proposed in literature. While KLE shows slight improvement, it still exhibits similar sensitivity issue since a new tuning parameter called temporal correlation range was introduced. Recently, additional effort was paid to illustrate the PCA-poisoning problem proposed in related work. To underline this problem, an evading strategy called BoiledFrog was proposed which adds a high fraction of outliers to the traffic. To defend against this, the authors employed a more robust version of PCA called PCA-GRID. While PCA-GRID shows performance improvement regarding the robustness to the outliers, it experiences a high sensitivity to the threshold estimate and the k-dimensional subspace that maximizes the dispersion of the data. The purpose of this work is to consider another technique to address the PCA poisoning problems to provide robust traffic anomaly detection: The Principal Component Pursuit.

5. Xiangliang Zhang, Wei Wang, "Self-adaptive Change Detection in Streaming Data with Non-stationary Distribution", 6th International Conference on Advanced Data Mining and Applications (**ADMA' 2010**), pp. 334-345, Springer, ChongQing, China, November 19-21, 2010 (Full paper).

   Abstract: Non-stationary distribution, in which the data distribution evolves over time, is a common issue in many application fields, e.g., intrusion detection and grid computing. Detecting the changes in massive streaming data with a non-stationary distribution helps to alarm the anomalies, to clean the noises, and to report the new patterns. In this paper, we employ a novel approach for detecting changes in streaming data with the purpose of improving the quality of modeling the data streams. Through observing the outliers, this approach of change detection uses a weighted standard deviation to monitor the evolution of the distribution of data streams. A cumulative statistical test, Page-Hinkley, is employed to collect the evidence of changes in distribution. The parameter used for reporting the changes is self-adaptively adjusted according to the distribution of data streams, rather than set by a fixed empirical value. The self-adaptability of the novel approach enhances the effectiveness of modeling data streams by timely catching the changes of distributions. We validated the approach on an online clustering framework with a benchmark KDDcup 1999 intrusion detection data set as well as with a real-world grid data set. The validation results demonstrate its better performance on achieving higher accuracy and lower percentage of outliers comparing to the other change detection approaches.

III -Attended Seminars, Workshops, and Conferences

1. The Secrets of Cryptography seminar in Luxembourg, Luxembourg, 28-29 June, 2010.

2. Summer School 2010: Verification Technology, Systems & Applications, Luxembourg, 6-10 September, 2010.
3. The 6th International Conference on Advanced Data Mining and Applications (ADMA2010), Chongqing, China, 19-21 November, 2010.
4. The Sixth International Conference on Information Systems Security (ICISS 2010), Gandhinagar, India, 15-19 December 2010.

IV – Research Exchange Programme (12 month scheme)

First Visit: Q2S Centre (the Centre for Quantifiable Quality of Service in Communication Systems), Norwegian University of Science and Technology (NTNU) (Mar- June 2010)
Contact Professor: Svein J. Knapskog (lastname@q2s.ntnu.no).

I decided to visit Q2S Centre for a longer time as I have stayed Q2S last year when I was an ERCIM fellow at NTNU. I have some collaboration work on network traffic analysis. During my visit, I have discussed with Atef Abdelkefi in network anomaly detection. This leads a paper accepted. I also collaborated with Xiangliang Zhang to work on developing data mining algorithms. I would say this visit is very productive.

Second Visit: CNR, Italy (September 2010)
Contact Professor: Dr. Fabio Martinelli (firstname.lastname@iit.cnr.it)

During my research exchange program at CNR, I visited the Security group. I have made a presentation entitled “Autonomic web attack detection” in the group seminar. I have discussed with Dr. Martinelli and Dr. Daniele Sgandurra for the related work and potential future collaboration.

Acknowledgements

First of all, I would like to thank Prof. Björn Ottersten for providing me the great opportunity to work at the SnT Centre (Interdisciplinary Centre of Security, Reliability and Trust) inaugurated in 2009. Under his support, I am also able to attend the ICISS 2010 conference during my ERCIM stay. He is so kind to extend my stay at SnT after the ERCIM fellowship. Many thanks go to Prof. Yves Le Traon. He has provided me several opportunities for funding applications. I have also collaborated with him in his SERVAL research group. In addition, I also thank Prof. Pascal Bouvry for his support for the application of Core project. I benefit a lot from the support of the professors and staff at SnT. Finally I would like to thank Ms. Emma Liere for her help in administrative matters during my ERCIM stay.