E-Mail Spam detection by Collaborative Reputation-Based Vector Space Model (CRVSM) and effective performance study

Masrath Parveen[1], Dr. Saurabh Pal[2], Dr. Venkateswara Rao CH[3]

[1]Research Scholar, Dept of CSE, V.B.S.Purvanchal University, Jaunpur
[2] Department of CSE, V.B.S.Purvanchal University, Jaunpur
[3]Department of CSE, Siddhartha Institute of Engineering and Technology, Hyderabad

Abstract: - In order to conduct an offence, malevolent attacks like spamming, phishing, or hacking are considered into cyber crime. The computer systems are deftly hacked and compromised, causing significant financial loss that might huge impact. Email spamming is the most well-known type of cyber attack since it uses up more cyber resources, such as memory, computing power, network bandwidth, traffic abuse, etc. Spam emails are mass-produced, unsolicited commercial emails that are sent for a variety of reasons. Studies show that more than 85% of today's email is spam. Researchers have come up with a number of strategies to control email spam, but some of them have been comprehensive or successful. The methods for detecting email spam that are currently in use have some major drawbacks. First of all, it has not been possible to efficiently separate spam emails from legitimate ones. This has increased the amount of false positives and false negatives, which has reduced the accuracy of detection. Second, when the volume of emails received rises, it takes longer to identify spam emails. Thirdly, because the detection filters are installed on the server, the server is overworked while handling large operations. Therefore, effective reaction mechanisms and efficient collaborative detection techniques are found for early, widespread identification and mitigation of spam emails and their source at the receiver side. User Authorization Phase, Feature Extraction Phase, Classification Phase, and Similarity Detection Phase are the four steps of an unique Probabilistic EShield Protocol (PEP) that provides Email Spam Detection with extra features. By evaluating the email content utilizing extra functions, PEP filters the spam email as well as the illegal sender of the incoming email. The results of experiments conducted for PEP demonstrate that PEP outperforms CRVSM in terms of detection accuracy, false positive and false negative rate reduction, and detection rate attainment. Therefore, three unique protocols have been presented in this study for email spam detection, offering cyber security in the cyber- space: Collaborative Reputation-Based Vector Space Model (CRVSM), Probabilistic EShield Protocol (PEP), and Optimized Feature Selection Protocol (OFSP).

Keywords: - CRVSM, mail spam, SDT, SDR, OFSP and PEP.

1. Introduction

Today, the spam of the mail can detected From 2010 until the present, filters against email spam relied on individual users to handle their incoming emails with complex features since email spam is known to be pervasive, repeated, and inescapable. Big Data's introduction enables ISPs to keep track of user inbox activity to determine if an incoming email is spam or not. Thus, Big Data functionality offers efficient and effective filtering.

Future email spamming may be impossible to stop, thus comprehensive response measurement is required to continuously monitor incoming emails and to evaluate recipients' reaction over time. Similar to this, several approaches were put into place on the server, however this places a significant pressure on the server to do a high number ofjobs in a short amount of time. It would be welcomed if anti-spam measures were implemented cooperatively at the recipient (Ashish Malviya et al. 2011,) so that the server could be relieved of its load and the detection could be completed quickly.

2. OBJECTIVE OF THE WORK

In order to detect and mitigate spam emails and their source as early as feasible at the receiver side using efficient algorithms, the work's goal is to build effective co-operative detection and mitigation methods. A thorough review of the literature on detecting email spam online allowed for the identification of issues and the formulation of goals. On the basis of accurately and plainly addressing these issues, the following objectives of the proposed work have been established:

1. To more accurately and with fewer false positives and false negatives, effectively separate spam emails from non-spam emails in cyberspace.
2. To shorten the time spent on spam detection by using cooperative detection.

To carry out a cooperative detection at the receiver side, lessening the strainon the server.

3. PROBLEM STATEMENT:

First, they were unable to properly distinguish between spam emails and non- spam emails, which is one of the fundamental weaknesses of the currently available methodologies in the field of email spam detection. As a result, there are more false positives and negatives and the detection accuracy is decreased. Second, as the volume of incoming emails rises, so does the detection time. Thirdly, server-side deployment of filters prevents the server from completing large jobs quickly.

The following difficulties with the spam detection techniques were able to be clearly and accurately defined thanks to the thorough literature review in the area of email spam detection in cyberspace:

First off, it has not been done effectively to separate spam emails from other communications in cyberspace. As a result, detection accuracy is decreased and there are more false positives and false negatives.

Second, the introduction of email sender authorisation was unsuccessful. As a result, there are now more people sending spam emails.

Thirdly, the time it takes to identify spam emails and their senders gets longer. Therefore, the time delay can be decreased more if collaborative detection is possible with more characteristics.

Fourth, because spam detection is handled on the server, the server is overworked. Implementing distributed spam detection at the receiver side with sophisticated response measurement for continuously evaluating incoming emails and for continua l feedback assessment of the receivers can solve this issue.

The following contributions were made:

- A Collaborative Reputation-Based Vector Space Model (CRVSM) has been designed and implemented. The performance of CRVSM has been analyzed using adequate experiments. This protocol performs accurate and effective classification of incoming emails in three phases: Feature Extraction, Similarity Detection and Collaborative Reputation Evaluation. CRVSM focuses on reputation-based detection of spam emails mainly at the receiver side in cyber space.

- A Probabilistic EShield Protocol (PEP) has been designed and implemented to perform email spam detection with additional features. The performance of PEP has been analyzed using adequate experiments. The PEP protocol performs Email Spam Detection in four phases: User Authorization, Feature Extraction, Classification and Similarity Detection. The two main tasks of PEP protocol are: First, it performs authorization of the received email to identify the unauthorized sender. Second, it analyzes the email contents to filter out the spam emails. This protocol is very effective, provides accurate detection and reduces the number of false positives and false negatives.

An Optimized Feature Selection Protocol (OFSP) has been designed and implemented as a hybrid rule-based approach that combines two well-known feature selection methods for email spam filtering. The OFSP protocol functions in four steps: Feature Selection, Normalization, Score Assignment and Optimal Feature Selection. This protocol is applicable for large datasets and achieves reduction of features optimally and efficiently with reasonable complexity. The performance of OFSP has been analyzed using adequate experiments. This protocol provides an optimal solution for email spam detection and outperforms CRVSM and PEP protocols.

4. EXPERIMENTAL RESULT OF CRVSM PROTOCOL

Java programming model has been used for the implementation and evaluation of the proposed CRVSM protocol. Experiments were conducted on a dataset that contains 1.4 million emails. The performance of CRVSM protocol has been analyzed with 5000, 10000, 15000, 20000 and 25000 emails with a spam ratio of 0.5%. The CRVSM protocol has been executed on a general processor computer (Intel(R) i5 processor) with 2.67 GHz and 8 GB RAM. The following metrics were used to evaluate the performance of CRVSM: The performance metrics of CRVSM protocol have been evaluated by varying percentage of collaborative reporters, the number of senders and the number of incoming emails from each sender. The percentage of collaborative reporters has been varied from 20% to 80% of total number of receivers of relevant email. The reporters work together in a collaborative fashion. The number of senders has been varied from 100 to 500 and the number of incoming email has been varied from 5000 to 25000. The CRVSM model is dynamic and is more effective over time.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

Experiment 1:

The experiment is conducted to study the FPR for 100 to 500 senders by varying the number of emails being sent as, 5000, 10000, 15000, 20000 and 25000 emails. Figure 1 shows the FPR for different number of email senders with varying number of emails.

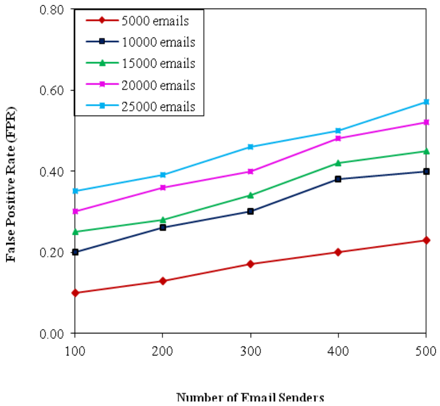


Figure 1 FPR for CRVSM Vs Number of Email Senders

Experiment 2:

The experiment is conducted to compare the FPR of CRVSM protocol with other four protocols viz. MLP, FEDM, PM and VSM for varying percentage of collaborative reporters. The experiment is conducted for 500 senders and 25000 emails. Figure 2 shows the comparison of FPR for CRVSM with four existing protocols for different percentage of collaborative reporters.

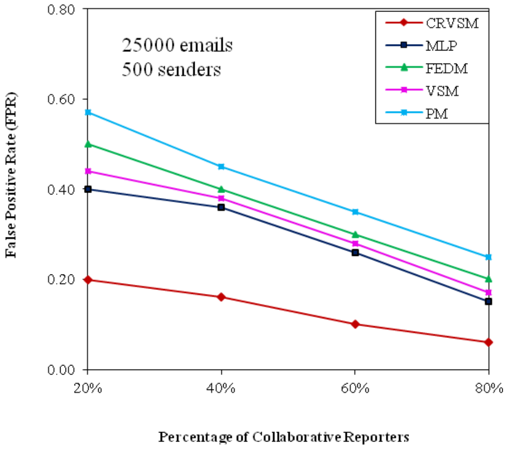
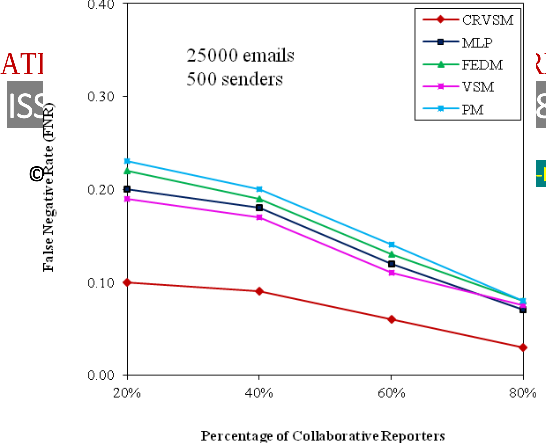


Figure 2 Comparison of FPR for CRVSM

It has been observed that as the number of collaborative reporters increases, the FPR of CRVSM decreases and provides better results than existing protocols. The results show that with 500 senders and 25000 emails being sent and for 20% of collaborative reporters, CRVSM generates a FPR of 0.20 and as the percentage of collaborative reporters increases (i.e., 80%), the FPR of CRVSM decreases to 0.06 and provides better results than existing protocols. CRVSM outperforms the other four protocols by generating less FPR.

Experiment 3



The experiment is conducted to study the FNR for 100 to 500 senders by varying the number of emails being sent as, 5000, 10000, 15000, 20000 and 25000 emails.
Figure 3 shows the FNR for different number of email senders with varying number of emails.

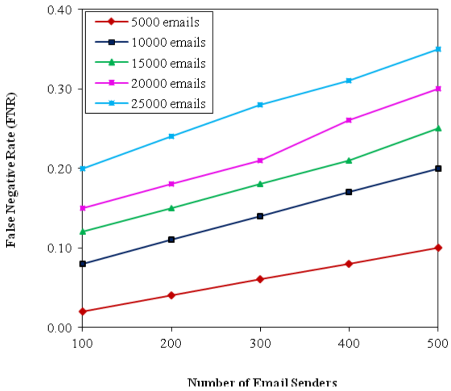


Figure 3 FNR for CRVSM Vs Number of Email Senders

It has been observed that with less number of senders, only less number of emails reach the receiver and hence the false negatives decreases and it increases as the number of senders increases with increasing number of emails. The increase in the number of false negatives increases the FPR value and vice versa.

Experiment 4

The experiment is conducted to compare the FNR of CRVSM protocol with other four protocols viz. MLP, FEDM, PM and VSM for varying percentage of collaborative reporters. The experiment is conducted for 500 senders and 25000 emails. Figure 4 shows the comparison of FNR for the five protocols for different percentage of collaborative reporters.

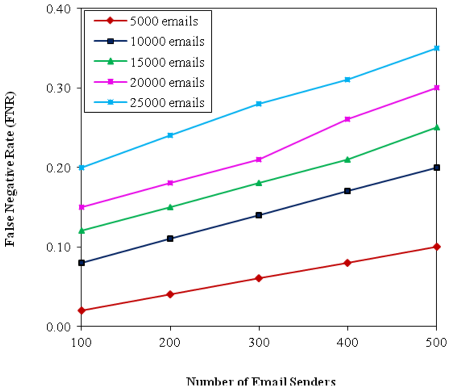


Figure 4 Comparison of FNR for CRVSM

It has been observed that as the number of collaborative reporters increases, the FNR of CRVSM decreases and provides better results than existing protocols. The results show that with 500 senders and 25000 emails being sent and for 20% of collaborative reporters, CRVSM generates a FNR of 0.10 and as the percentage of collaborative reporters increases (i.e., 80%), the FNR of CRVSM decreases to 0.03 and provides better results than existing protocols. CRVSM outperforms the other four protocols by generating less FNR.

Experiment 5

The experiment is conducted to compare the overall throughput of CRVSM with other four protocols viz. MLP, FEDM, PM and VSM by varying the number of emails with a bandwidth value of 3 Mbps for 500 senders with 80% of collaborative reporters. Figure 5 shows the overall throughput for the five protocols for varying number of emails with a bandwidth value of 3 Mbps. It has been observed that CRVSM outperforms the other four protocols by efficiently utilizing the network bandwidth values and thus achieves good overall throughput.

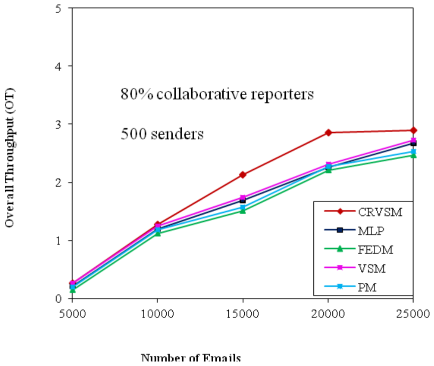


Figure 5 Comparison of Overall Throughput for CRVSM

6. COMPLEXITY ANALYSIS OF CRVSM

The time complexity has been analyzed $2T$ mathematically for CRVSM model. The CRVSM model performs feature extraction in T^2 time step followed by similarity detection in time step and collaborative reputation evaluation in time step for all the email being received at the receiver. Therefore, the total time complexity of CRVSM has been formulated in (4.7) as,

$$TC_{CRVSM} \leq T^2 + 3T \qquad T \qquad (4.7)$$

VSM requires time step for cluster generation and T^2 time step for reduced dataset generation. Therefore, the total time complexity of VSM has been formulated as, $TC_{VSM} \leq T^2 + T$.

FEDM requires time step for construction of feature set, T^2 time step for reduced feature set generation and T^2 time step for classification using Cluster-based Classification. Therefore, the total time complexity of FEDM has been formulated as, $TC_{FEDM} \leq T + 2T^2$.

PM requires T^2 time step for generation of decision trees for classification purposes and time step for filtering spam email at each stage (i.e., four pipeline stages) of pipeline. Therefore, the total time complexity of PM has been formulated as, $TC_{PM} \leq T^2 + 4T$.

MLP requires time step for T clustering the email into different layers and T^2 time step for activation of neurons at a layer. Therefore, the total time complexity of MLP has been formulated as, $TC_{MLP} \leq T^2 + T$

The presented and evaluated the experimental results and performance results of the novel CRVSM protocol. The experimental results show that the novel CRVSM protocol outperforms the other protocols like MLP, FEDM, VSM and PM. The CRVSM protocol achieves accurate detection of spam emails by reducing the false positive rate, false negative rate and thereby increasing the detection accuracy to a greater degree. The CRVSM protocol achieves timely detection by reducing the spam detection time. Moreover, the CRVSM protocol proves its efficiency by achieving good spam detection rate. The CRVSM protocol provides guaranteed system performance by achieving good network service ratio and overall throughput.

7. CONCLUSION AND FUTURE SCOPE

Email Spamming is the most recognized form of cyber attack that causes heavy financial loss in trillions in the cyber-space. Email spammers compromise the computer systems and exploit their resources by transmitting spam emails massively. Researchers are proposing several approaches against email spamming but they cannot provide a complete and effective solution.

In this paper, the recently proposed research approaches against email spamming, their descriptions, advantages and disadvantages have been surveyed and clearly presented. The survey on these research approaches in detecting spam emails enabled to clearly define the problems that exist in the cyber space on Email Spam Detection. The gaps such as, inefficiency in isolating spam emails from non-spam ones, increased detection delay, higher false alarms and overhead in complexity that have made the system less efficient have been identified through this survey. These gaps can be bridged through efficient email spam detection protocols.

Three novel email spam detection protocols viz. Collaborative Reputation- Based Vector Space Model (CRVSM), Probabilistic EShield Protocol (PEP) and Optimized Feature Selection Protocol (OFSP) have been proposed to bridge these gaps analyzed in the existing approaches. The proposed protocols increase the system performance and improve the system efficiency.

REFERENCES

1. Aakash Atul Alurkar; Sourabh Bharat Ranade; Shreeya Vijay Joshi; Siddhesh Sanjay Ranade, Piyush A Sonewar, Parikshit N Mahalle & Arvind V Deshpande 2017, 'A proposed data science approach for email spam classification using machine learning techniques', Internet of Things Business Models, Users, and Networks, pp. 1-5.

2. Abdelrahman AlMahmoud, Ernesto Damiani, Hadi Otrok & Yousof Al-Hammadi 2017, 'Spamdoop: A privacy-preserving Big Dataplatform for collaborative spam detection', IEEE Transactions on Big Data, issue 99.

3. Alan Gray & Mads Haahr 2004, 'Personalised, Collaborative Spam Filtering', in proceedings of *the First Conference on Email and Anti- Spam (CEAS), Mountain View, CA, USA, July-August, under grant no. CFTD/03/219.*

4. *Ali Shafigh Aski & Navid Khalilzadeh Sourati 2016, 'Proposed efficient algorithm to filter spam using machine learning Techniques', Pacific Science Review A: Natural Science and Engineering, vol. 18, issue 2, pp. 145-149.*

5. *Amani Mobarak & AlMadahkah 2016, 'Big Data In computer Cyber Security Systems', IJCSNS International Journal of Computer Science and Network Security, vol. 16, no. 4, pp. 56-65.*

6. *Andreas GK Janecek, Wilfried N Gansterer & Ashwin Kumar, K 2008, 'Multi-Level Reputation-Based Greylisting', in proc. of Third International Conference on Availability, Reliability and Security ARES 08, 4-7 March 2008, Barcelona, Spain.*

7. *Andrej Bratko, Gordon V Cormack, Bogdan Filipi˘c, Thomas R, Lynam & Blaz Zupan 2006, 'Spam Filtering Using Statistical Data Compression Models', Journal of Machine Learning Research, Cheriton School of Computer Science, University of Waterloo, Ontario 3G1, Canada, vol. 7, pp. 2673-2698.*

8. *Arunabha Mukhopadhyay, Samir Chatterjee, Debashis Saha, Ambuj Mahanti & Samir K Sadhukhan 2013, 'Cyber-risk decision models: To insure IT or not?', Decision Support Systems, http://dx.doi.org/ 10.1016/j.dss.2013.04.004, Volume 56, December 2013, pp. 11-26.*

9. *Ashish Malviya, Glenn A Fink, Landon Sego & Barbara Endicott- Popovsky 2011, 'Situational Awareness as a Measure of Performance in Cyber Security Collaborative Work', Eighth International Conference on Information Technology: New Generations, pp. 937- 942.*

10. *Barry Leiba, Joel Ossher, Rajan, Richard Segal, VT & Mark Wegman 2005, 'SMTP Path Analysis', CEAS 2005, Second Conference onEmail and Anti-Spam, Stanford University, California, USA.*

11. *Bekkerman, R, McCallum, A & Huang G 2004, 'Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora', Computer Science Department Faculty Publication Series, University of Massachusetts – Amherst Corpora, Center for Intelligent Information Retrieval, Technical Report IR.*

12. *Ben Medlock 2006, 'An Adaptive Approach to Spam Filtering on a New Corpus', CEAS 2006, Mountain View, California, USA.*

13. *Benjamin J Kuipers, Alex X Liu Aashin Gautam & Mohamed GGouda 2005, 'Zmail : Zero-Sum Free Market Control of Spam', in proc. of 25th IEEE International Conference on Distributed Computing Systems Workshops, 6-10 June 2005, Columbus, OH, USA.*

14. *Bin Srinidhi, Jia Yan & Giri Kumar Tayi 2015, 'Allocation of resources to cyber-security: The effect of misalignment of interestbetween managers and investors', Decision Support Systems, July 2015, vol. 75, pp. 49-62.*

15. *Byrnea, DJ, David Morganb, Kymie Tana, Bryan Johnsona & Chris Dorrosa 2014, 'Cyber Defense of Space-Based Assets: Verifying and Validating Defensive Designs and Implementations', Conference on Systems Engineering Research (CSER 2014), Science Direct, vol. 28, pp. 522-530.*

16. *Carlos Laorden, Borja Sanz, Igor Santos, Patxi Galán-García & Pablo G Bringas 2013, 'Collective classification for spam filtering', Logic Journal of the IGPL, vol. 21, issue 4, pp. 540-548.*

17. *Carlos Laorden, Xabier Ugarte-Pedrero, Igor Santos, Borja Sanz, Javier Nieves & Pablo G Bringas 2014, 'Study on the effectiveness of anomaly detection for spam filtering', Information Sciences, http://dx.doi.org/10.1016/j.ins.2014.02.114, Science Direct, vol. 27, pp. 421- 444.*

18. *Cherry, S 2006, 'Cisco and Yahoo's plan to Damn Spam', IEEE Spectrum, vol. 43, issue 1, pp. 51-51.*

19. *Chi-Yao Tseng, Pin-Chieh Sung & Ming-Syan Chen 2011, 'Cosdes: A Collaborative Spam Detection System with a Novel E-Mail Abstraction Scheme', IEEE Transactions on Knowledge and Data Engineering, vol. 23, issue 5, pp. 669-682.*

20. *Christina, V, Karpagavalli, S & Suganya, G 2010, 'A Study on Email Spam Filtering Techniques', International Journal of Computer Applications, ISSN : 0975 – 8887, vol. 12, no. 1.*

21. *Cosima Rughinis & Razvan Rughinis 2014, 'Nothing ventured, nothing gained. Profiles of online activity, cyber-crime exposure, and security measures of end-users in European Union', Computers & Security, ScienceDirect, http://dx.doi.org/10.1016/j.cose.2014.03.008,vol. 43, pp. 111-125.*

22. *Damiani, E, De Capitani di Vimercati, S, Paraboschi, S & Samarati, P 2004, 'P2P-based collaborative spam detection and filtering', Fourth International Conference on Peer-to-Peer Computing,*

DOI: 10.109/ PTP.2004.1334945, August 2004, pp. 176-183.

23. *Danny Bradbury 2014, 'Can we make email secure?', Network Security, vol. 2014, issue 3, pp. 13-16.*

24. *De Wang, Danesh Irani & Calton Pu 2013, 'A study on evolution of email spam over fifteen years', 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing,pp. 1-10.*

25. *Debin, G, Michael, KR & Dawn, S 2009, 'Beyond Output Voting: Detecting Compromised Replicas Using HMM-Based Behavioral Distance', IEEE Transactions on Dependable and Secure Computing, vol. 6, issue 2, pp. 96-110.*

26. *Derek E Bambauer, John G Palfrey, Jr. & David E Abrams 2005, 'A Comparative Analysis of Spam Laws: The quest for a model Law', International Telecommunication Union ITU WSIS Thematic Meeting on Cyber Security, Berkman Center for Internet & Society Harvard Law School, Geneva, Switzerland.*

27. *Dhanalakshmi Ranganayakulu & Chellappan, C 2013, 'Detecting Malicious URLs in Email – An Implementation', AASRI Conference on Intelligent Systems and Control, Elsevier B.V., Doi: 10.1016/ j.aasri. 2013. 10.020, vol. 4, pp. 125-131.*

28. *Drucker, H. Donghui Wu & Vapnik, VN 1999, 'Support vector machines for spam categorization', IEEE Transactions on Neural Networks, vol. 10, issue 5, pp. 1048-1054.*

29. *Enrico Blanzieri & Anton Bryl 2008, 'A Survey of learning based Techniques of Email spam filtering', Information Engineering and Computer science Department, University of Trento, Italy, vol. 29, issue 1, pp. 63-92.*

30. *Eric Filiol 2012, 'New Trends in Security Evaluation of Bayesian Network-based Malware Detection Models', 45th Hawaii InternationalConference on System Sciences, Operational Cryptology and VirologyLaboratory, ESIEA Group / ESIEA Research, DOI 10.1109/HICSS. 2012.450, pp. 5574-5583.*