3

HIGH PERFORMANCE COMPUTING & CYBER SECURTY

In contemporary research computation is the main pillar of scientific discovery, which provides an in-expensive way to achieve high science, complementing theory, experiment and observation. The capability and credibility of a scientific organization is judged by the computational facility the researchers have access to. CSIR-4PI provides state-of-the-art High Performance Computing facility to the computational scientists and researchers across CSIR to address Grand Challenge problems in their frontier areas of science and engineering. The facility at CSIR-4PI is a centralized High Performance Computing facility. It is one of the top supercomputers of the country and provides multiple architectures suitable for many different domain specific applications. All the CSIR laboratories access the facility through the high speed National Knowledge Network. In addition to providing the HPC facility, the group is also involved in research on cyber security which is an important area for the future. Extending the research carried out under the 12th Five Year Plan of CSIR the team have got a mission mode project on Intelligent Systems, in which the team is concentrating on the security and privacy issues in connected vehicles and biometric based transactions.

Inside

- Darknet traffic analysis for cyber security inference
- Network traffic processing and analysis using multi-node Apache Spark and YARN environment
- Deep learning framework for short-term wind speed forecast
- Design of secure cryptographic hash function using soft computing techniques
- High Performance Computing

3.1 Darknet traffic analysis for cyber security inference

Darknet refers to the range of unused IP addresses i.e. the IP addresses which do not have any host attached to them. The traffic observed in darknet is often referred as Internet Background Radiation (IBR) or unsolicited network traffic. IBR is a potential source for inferering various malicious activities in the cyberspace and revealing associated trends.

Clustering techniques was the first choice to find distinct attack patterns in the unsolicited TCP network traffic. The IP header information was transformed to AGM (AGgregate and Mode) format [22 tuple] which was further converted to numerical AGM in order to apply the clustering techniques. Each AGM tuple corresponds to a distinct source IP address. Each AGM has the traffic information of the packets sent by that particular source IP. The numerical AGM format uses the parameters: destination IP, source port, destination port, TTL, flag and packet length. The observation time used to calculate each AGM was 24 hours. The data was preprocessed and reduced to three dimensions for efficiency by using Principal Component Analysis. Then Mean Shift Clustering, a hierarchical clustering algorithm, was applied to the preprocessed data. After applying these set of operations to a month's data (July 2017), we could find the traces of Mirai bot, the well-known IoT bot on the Internet, and malicous attempts like SQL injection and remote desktop attacks.

Figure 3.1. Post-clustering patterns in the IBR traffic

Figure 3.1 shows the top four clusters, in terms of number of AGM present in the cluster, after applying the clustering technique to 3rd July 2017 data. In this plot, the red cluster largely corresponds to mirai traffic, the green is SQL injection traces and blue is attempts towards remote desktop attack.

3.2 Network traffic processing and analysis using multi-node Apache SPARK and YARN environment

An experimental environment has been setup with Apache SPARK on top of Apache Hadoop YARN (Yet Another Resource Negotiator) cluster in the cyber security test-bed. The setup was used to analyze 85GB of network packet, collected over a period of four months. We are able to analyze 85GB of data in just 78 seconds, using 32 node (256 cores) SPARK cluster. This would otherwise take around 30-40 minutes in traditional processing systems. Best results were achieved by allotting 7 out of 8 cores of CPU per node, compared to 5 and 8 cores per node.

Table 3.1 Time Taken for Processing of Four Months Data (85GB), using 7 CPU-Cores Per Node

No. of nodes (cores)	Time taken in seconds
1 node (7 cores)	971
2 nodes (14 cores)	592
4 nodes (28 cores)	360
8 nodes (56 cores)	218
16 nodes (112 cores)	131
32 nodes (224 cores)	**78**

The results show that significant reduction in the processing time can be achieved by using emerging data analytics tools, and there is a near-linear scalability.

3.2 Deep learning framework for short-term wind speed forecast

Wind power is one of the most popular modern and sustainable renewable energy sources among other renewable energy sources like hydropower, bio-mas energy, ocean energy etc. Renewable energy sources have major impact not only on economic development but also on environmental and social development. In power and energy sector, electric power generation from wind energy receives a huge success as it is renewable and pollution free. Sustain wind is required to generate wind power energy. It helps in maximizing the energy protection and minimizing the operational cost. Accurate wind speed prediction plays an important role in wind energy generation. Hence an accurate prediction of wind speed and its frequency distribution help in calculating amount of electric power generation. But accurate wind speed prediction is quite challenging as wind has intermittent and stochastic behavior.

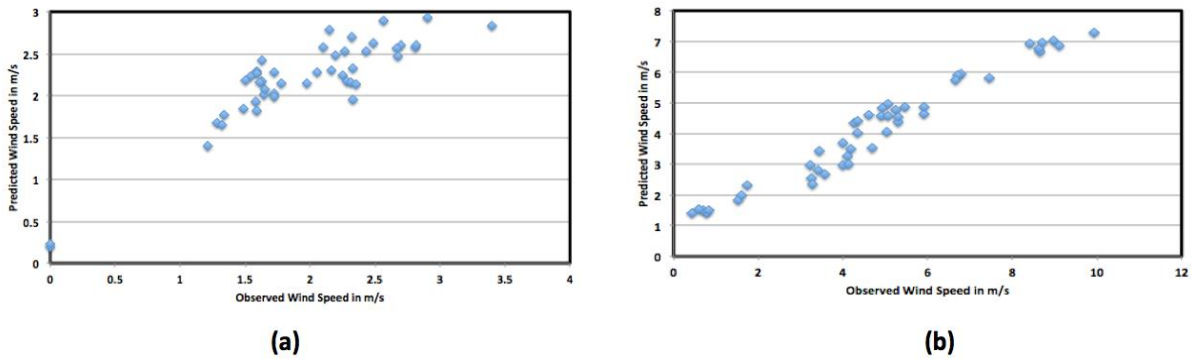

Figure 3.2 Weekly averaged predicted wind speed (m/s) vs observed wind speed (m/s) for (a) Delhi (b) Bangalore towers

We have used a modified framework using Deep Learning Technique for improving the accuracy of wind speed prediction, and have achieved substantial improvement in terms of accuracy. For our studies we have considered data from two towers established by CSIR, one in New Delhi and the other at Bengaluru. The availability of data for these towers is for the period 2009-2014 (5

years). We have used four meteorological parameters i.e. temperature (t), pressure (p), humidity (h) and wind speed (v) available at 30 minutes averaged interval. The observed values are from sensors mounted at 20-meter height from the ground level. The number of hidden neurons, hidden layers and number of epoch are tuned continuously in order to arrive at the least error possible. We have analyzed the results using different statistical methods. Figure 3.2 shows the plot of weekly averaged wind speed, with the observed wind speed in *x-axes* and the predicted wind speed in the *y-axes*. The result shows a good correlation between the observed and the predicted at both the locations.

3.3 Design of secure cryptographic hash function using soft computing techniques

Data integrity is a crucial part of any secure system. Cryptographic hash functions serve as a basic building block of information security for ensuring data integrity and data origin authentication. They are used in numerous security applications such as digital signature schemes, construction of MAC and random number generation. A hash function takes an arbitrary amount of input and produces an output of fixed size. Many of the widely used cryptographic MD-5 and SHA-1 hash functions have been shown to be vulnerable to attacks. The non-linear behavior of the neural network model, which takes multiple inputs to produce single output, makes it a perfect entrant for cryptographic hash design. We have designed a cryptographic hash function using a multi layer Tree Parity Machine neural network.

Table 3.2 Comparison of the proposed methods other standard Hash function

Algorithm and variant	Output size (bits)	Rounds	Operations
MD5	128	64	And, Xor, Rot,Add (mod 2^{32}), Or
SHA-1	160	80	And, Xor, Rot,Add (mod 2^{32}), Or
SHA-256	256	64	And, Xor, Rot, Add (mod 2^{32}), Or,Shr
SHA-512	512	80	And, Xor, Rot,Add (mod 2^{64}), Or,Shr
Proposed Hash	512	16	Mul, Add, And

Table 3.2, shows the comparison of the proposed method to the other popular hash functions. It can be observed that, in terms of number of rounds and as well as the number of logical units to be implemented, our proposed method is better than the other hash function currently in use. Although in our simulations we have considered 512 bit message blocks, our algorithm can be used flexibly to generate a hash function of arbitrary length. Simulations show that this hash function satisfies the security requirements of confusion, diffusion, and collision attack

3.4 High Performance Computing

High Performance Computing (HPC) is one of the flagship activities of CSIR-4PI. This facility provides the necessary computational support to the whole of CSIR (more than 200 scientists) over the National Knowledge Network. The computational facility, “Ananta” supercomputer, with peak theoretical computing capability of 360TF and a sustained High Performance LINPACK

(HPL) of 334TF is the main lifeline of the computational scientists of CSIR over the last five years and is the largest supercomputer of CSIR. In addition the center host an ALTIX-ICE medium range HPC along with a hierarchical storage infrastructure.

Figure 3.3 CSIR centralized 360TF High Performance Computing Facility.

The “Ananta” supercomputer, is very heavily used and is currently listed as the 8th fastest system in the country. Ananta, is a cluster of 1088 computing nodes, distributed over 17 racks. Each node is having 4GB memory per core, which results in of about 68TB of distributed memory for the total system. The inter-node communication is powered by high speed FDR infiniband (providing a dedicated 56 Gbps interconnect bandwidth) in a FAT tree topology. The supercomputer also has an online storage using LUSTRE parallel file system of about 2.1 Peta Byte and is capable of providing a minimum of 20 Gbps simultaneous read and write performance. PBSPro workload manager ensures efficient usage of the system. To store and archive the results, an archival system based on a high performance 3-tired storage SAN (Storage Area Network) is established and upgraded regularly to support the growing need of storage.

Ananta is located in a Tier-3 equivalent state-of-the-art data center efficiently supported by a state-of-the-art energy farm. The most noteworthy component of the datacenter is the water based cooling mechanism called Rear Door Heat Exchangers (RDHX). Due to this the datacenter is one of the high density and high power efficient datacenter (Power Usage Efficiency (PUE) of less than 1.5) in the country. An energy farm consisting of two numbers of redundant compact substations of 1.25MVA each and for ensuring 24X7 power supply to the datacenter three numbers of diesel generators, an underground diesel yard (more than 15000 liters), three numbers of UPS with battery backup supports the data centre.