**pSigene: Webcrawling to Generalize SQLi Signatures**

Gaspar Modelo-Howard, Fahad Arshad, Chris Gutierrez, Saurabh Bagchi, Yuan Qi

**Problem Statement**
- Misuse-based detection systems use signatures of attacks to detect malicious activity, which require to be continuously updated
- Current approach to create and update signatures is manual
- Signatures to improve detection systems, are necessary to complement prevention mechanisms

**Specific Goals**
- Define process to **automatically generate detection signatures** by performing data mining on attack samples
- Create **generalized signatures**, matching for attacks and its variations

**Proposed Solution**
- Framework for the automatic creation of generalized signatures represented as collection of regular expressions, by applying a sequence of two data mining techniques to a corpus of attack samples
- Solution suggests number of signatures necessary to detect attacks, while helping reduce size of signatures
- We demonstrate our solution specifically with SQL injection (SQLi) attacks, which have been very dominant in the last couple of years

**Experimental Results**
1. Collected over 30k SQLi attacks samples from 2 cybersecurity portals
2. Characterized each sample using set of 159 features from 3 sources: SQL reserved words, NIDS/WAF SQLi signatures, and SQLi reference documents
3. Performed a 2-way hierarchical agglomerative clustering (HAC) algorithm, using UPGMA and Euclidean distance to produce 9 bi-clusters

**pSigene Example: Signature 6**
```
"[\\?&]\[^[^a-zA-Z&]\]+?"  "\([\?&]\[^[^a-zA-Z&]\]+?\)\?;"
```
- Signatures were implemented in Bro NIDS with function that returned number of times a feature was found in a HTTP request (count_all(fi,j, reqHTTP))

**Evaluation**
- Test Set: 1.4M (benign) and 7.2k (malicious) HTTP GET requests
- ROC Curves for each of the pSigene Generalized Signatures

<table>
<thead>
<tr>
<th>RULES</th>
<th>TPR(%)</th>
<th>FPR(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bro</td>
<td>73.23</td>
<td>0.00</td>
</tr>
<tr>
<td>Snort – Emerging Threats</td>
<td>79.55</td>
<td>0.1742</td>
</tr>
<tr>
<td>ModSecurity</td>
<td>96.07</td>
<td>0.0515</td>
</tr>
<tr>
<td>pSiGene (9 rules)</td>
<td>86.53</td>
<td>0.037</td>
</tr>
<tr>
<td>pSiGene (7 rules)</td>
<td>82.72</td>
<td>0.016</td>
</tr>
</tbody>
</table>