**Motivation**
- Programs hold a lot of rules
- Source code contains useful information for inferring these rules
- Data mining is a suitable technique to extract information from large software source code

**Projects**
- CP-Miner
  - Detecting copy-paste code and related bugs
  - Published in OSDI’04
- PR-Miner
  - Extracting programming rules and detecting violations
  - Published in ESEC/FSE’05

**Overview of CP-Miner**
- Copy-pasting is common in large software
- Copy-pasted code is error prone
- CP-Miner
  - Can efficiently identify copy-paste in large software
  - Can identify copy-pasted code with modifications
  - Can detect copy-paste related bugs

  ```c
  for (i=0; i<n; i++) {
    total[i].adr = list[i].adr;
    total[i].size = list[i].size;
    total[i].more = &total[i+1];
  }
  ```

  Example: A copy-paste related bug detected by CP-Miner in Linux 2.6.11.

**How Does CP-Miner Work?**
- Identifying Copy-Pasted Code
  - Characteristics of copy-pasted code
    - Code based
    - Appears for at least twice
    - May be not exactly the same
  - Frequent sequence mining

**Results of CP-Miner**
- Bugs detected by CP-Miner were reported by us and are now fixed afterwards
  - Linux: 4.4 M, 22.3%, 20 min, 49
  - FreeBSD: 3.3 M, 20.4%, 20 min, 31
  - Apache: 224 K, 17.7%, 15 sec, 5
  - PostgreSQL: 458 K, 22.2%, 38 sec, 2

**Overview of PR-Miner**
- Programs follow many programming rules
- Most rules are implicit and undocumented
- Many rules can be very complex
- Violations lead to errors

**How Does PR-Miner Work?**
- Extracting programming rules
  - Idea: finding association among elements that are frequently used together in source code
  - Frequent itemset mining

<table>
<thead>
<tr>
<th>Function definition:</th>
<th>Identifiers:</th>
<th>Tokens:</th>
</tr>
</thead>
<tbody>
<tr>
<td>int twa_probe(struct pci_dev *dev, ...) {</td>
<td>Scsi_Host</td>
<td>92</td>
</tr>
<tr>
<td>struct Scsi_Host *host = NULL;</td>
<td>host_alloc</td>
<td>39</td>
</tr>
<tr>
<td>host = host_alloc(...);</td>
<td>add_host</td>
<td>68</td>
</tr>
<tr>
<td>add_host(host, &amp;dev-&gt;dev);</td>
<td>pol_dev</td>
<td>56</td>
</tr>
<tr>
<td>scan_host(host);</td>
<td>scan_host</td>
<td>36</td>
</tr>
</tbody>
</table>

Example: Violations detected in Linux 2.6.11.

**Results of PR-Miner**
- Rules extracted and bugs detected
  - 88-97% rules involve variables
  - Manually inspected the top 60 violations
  - Linux: 32,283, 42 min, 16
  - PostgreSQL: 6,128, 5 min, 6
  - Apache: 283, 1 min