Introduction

Detecting Synchronized Behavior. To maximize profits, fraudsters reuse different resources (e.g., fake accounts, IP addresses, and device IDs) over multiple frauds.

Challenges.
- Search-based dense block detection methods are not resistant to noise.
- Tensor decomposition methods tend to miss small fraud groups.
- Semi-supervised fraud detection methods rely on labels difficult to obtain.
- Large node weights. For example, a fraud user reusing the IPs to review fraud products has a large node weight.
- Large group size. Many users in one fraud.

Methods

Step1: Building ISG

Information Sharing Graph (ISG).
- The probability an entry has a at dimension $A_k$: $p^k(a) = Pr(t(A_k) = a)$.
- The self information of the event that $u_i$ and $u_j$ share a at dimension $A_k$ $I_{ij}^k(a) = \log(\frac{1}{p^k(a)})^2$.

Edge weight: $S_{ij}$: the suspicious level between entities' sharing. We use the pairwise value sharing between $u_i$ and $u_j$ across all dimensions $S_{ij} = \sum_{k=1}^{N} \sum_{a_{ij}} I_{ij}^k(a)$.

The self information of the event that $u_i$ uses a at dimension $A_k$ for m times $I_{i}^k(a) = \log(\frac{1}{p^k(a)})^m$.

Node weight: $S_i$: the suspicious level of the node. We use the self-value sharing for $u_i$ across all dimensions $S_i = \sum_{t=1}^{N} \sum_{a_{i}} I_{i}^k(a)$.

Subgraph Formed by Fraud Groups on ISG.
- Large node weights. For example, a fraud user reusing the IPs to review fraud products has a large node weight.
- Large edge weights. For example, two fraud users share the same IP and review the same unpopular product on Amazon. Because this sharing event has a low probability and high information, the edge weight is large.
- Large group size. Many users in one fraud.

Step2: D-Spot

Graph Partition. Delete low-weight edges and partition the ISG into multiple connected components $\mathcal{G}_1, \mathcal{G}_2, \ldots$. So the later computation could run in parallel.

Finding One Dense Subgraph from One Graph Partition.
- Input: one graph partition $\mathcal{G}$ of ISG.
- Objective: finding a subgraph $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$ on $\mathcal{G}$ that maximizes the suspiciousness density

$$F_{\mathcal{G}} = \frac{\sum_{u \in \mathcal{V}} S_u + \sum_{u \in \mathcal{E}} S_{ij}}{|\mathcal{V}|}$$

Algorithm: In each iteration, we delete a set of nodes that leads the density $F$ decreases least from the current node set. Finally, the algorithm returns the node set that maximizes $F$.

Finding Multiple Dense Subgraphs in Parallel. In each graph partition $\mathcal{G}_1, \mathcal{G}_2, \ldots$, find a single dense subgraph. Then we delete it and find the next dense subgraph.

Advantages
- Two-approximation guarantee of D-Spot.
- By removing a set of nodes at once, we reduce the number of iterations.
- Low computation complexity. $O(|\mathcal{V}|^2 + |\mathcal{E}|)$ for each graph partition.
- Robustness to noisy features.

Evaluation

Accurate Fraud User Detection. Three restaurant review datasets from Yelp.

<table>
<thead>
<tr>
<th>AUC</th>
<th>YelpChi</th>
<th>YelpNYU</th>
<th>YelpZip</th>
</tr>
</thead>
<tbody>
<tr>
<td>67K Entries</td>
<td>359K Entries</td>
<td>1.14M Entries</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>YelpChi</th>
<th>YelpNYU</th>
<th>YelpZip</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-Zoom[1]</td>
<td>0.9831</td>
<td>0.9451</td>
<td>0.9426</td>
<td></td>
</tr>
<tr>
<td>M-Biz[1]</td>
<td>0.9831</td>
<td>0.9345</td>
<td>0.9403</td>
<td></td>
</tr>
<tr>
<td>D-Cube[2]</td>
<td>0.9810</td>
<td>0.9223</td>
<td>0.9376</td>
<td></td>
</tr>
<tr>
<td>ISG+D-Spot</td>
<td>0.9875</td>
<td>0.9546</td>
<td>0.9529</td>
<td></td>
</tr>
<tr>
<td>ISG+D-Spot</td>
<td>0.9875</td>
<td>0.9546</td>
<td>0.9529</td>
<td></td>
</tr>
</tbody>
</table>

Robustness to Noisy Features. Registration information of 16,156 normal users and 9,961 fraud users. 'C' = 'crucial feature' and 'N' = 'noisy feature'.

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>1C</th>
<th>2C</th>
<th>2C+1N</th>
<th>2C+2N</th>
<th>2C+3N</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-Zoom[1]</td>
<td>0.7676</td>
<td>0.8880</td>
<td>0.8827</td>
<td>0.8744</td>
<td>0.8439</td>
<td></td>
</tr>
<tr>
<td>M-Biz[1]</td>
<td>0.7677</td>
<td>0.8842</td>
<td>0.8827</td>
<td>0.8744</td>
<td>0.8439</td>
<td></td>
</tr>
<tr>
<td>D-Cube[2]</td>
<td>0.7522</td>
<td>0.9201</td>
<td>0.8586</td>
<td>0.8312</td>
<td>0.7987</td>
<td></td>
</tr>
<tr>
<td>ISG+D-Spot</td>
<td>0.7699</td>
<td>0.9946</td>
<td>0.9935</td>
<td>0.9917</td>
<td>0.9859</td>
<td></td>
</tr>
</tbody>
</table>

High Scalability. Near-linear time with respect to the number of entries on three Amazon review datasets.

References