Content Acquisition Optimization
Yahoo Webmessenger

- Update data sent to individuals logged into Yahoo’s Instant Messenger service online
  - Online contact status, unread emails in Yahoo inbox
  - Usually small sessions (2-4kB)
- Sporadic collection (30,000 – 60,000 sessions per day)
- Intermittent bursts of collection against contacts of targets
  - Large numbers of sessions (20,000+) against a single targeted selector
  - Not collected against the target (online presence/unread email from target)
  - No owner attribution (metadata value limited to fact-of comms for emails, online presence events for buddies)
- Over a dozen selectors detasked in two weeks
  - Because a target’s contact was using/idling on Yahoo Webmessenger
  - Several very timely selectors (Libyan transition, Greek financial related)
Address Books

- Email address books for most major webmail are collected as stand-alone sessions (no content present*)
- Address books are repetitive, large, and metadata-rich
- Data is stored multiple times (MARINA/MAINWAY, PINWALE, CLOUDs)
- Fewer and fewer address books attributable to users, targets
- Address books account for ~22% of SSO’s major accesses (up from ~12% in August)

<table>
<thead>
<tr>
<th>Access (10 Jan 12)</th>
<th>Total Sessions</th>
<th>Address Books</th>
</tr>
</thead>
<tbody>
<tr>
<td>US-3171</td>
<td>1488453</td>
<td>237067 (16% of traffic)</td>
</tr>
<tr>
<td>DS-200B</td>
<td>938378</td>
<td>311113 (33% of traffic)</td>
</tr>
<tr>
<td>US-3261</td>
<td>94132</td>
<td>2477 (3% of traffic)</td>
</tr>
<tr>
<td>US-3145</td>
<td>177663</td>
<td>29336 (16% of traffic)</td>
</tr>
<tr>
<td>US-3180</td>
<td>269794</td>
<td>40409 (15% of traffic)</td>
</tr>
<tr>
<td>US-3180 (16 Dec 11)</td>
<td>289318</td>
<td>91964 (32% of traffic)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3257738</td>
<td>712366 (22% of traffic)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Provider</th>
<th>Collected</th>
<th>Attributed</th>
<th>Attributed%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo</td>
<td>444743</td>
<td>11009</td>
<td>2.48%</td>
</tr>
<tr>
<td>Hotmail</td>
<td>105068</td>
<td>1115</td>
<td>1.06%</td>
</tr>
<tr>
<td>Gmail</td>
<td>33697</td>
<td>2350</td>
<td>6.97%</td>
</tr>
<tr>
<td>Facebook</td>
<td>82857</td>
<td>79437</td>
<td>95.87%</td>
</tr>
<tr>
<td>Other</td>
<td>22881</td>
<td>1175</td>
<td>5.14%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>689246</td>
<td>95086</td>
<td>13.80%</td>
</tr>
</tbody>
</table>
Buddy Lists, Inboxes

• Unlike address books, frequently contain content data
  – Offline messages, buddy icon updates, other data included
  – Webmail inboxes increasingly include email content
  – Most collection is due to the presence of a target on a buddy list where the communication is not to, from, or about that target

• NSA collects, on a representative day, ~ 500,000 buddylists and inboxes
  – More than 90% collected because tasked selectors identified only as contacts (not communicant, content, or owner)

• Identifying buddylists and inboxes without content (or without useful content) an ongoing challenge
Scenario: @yahoo.com

- Sep 2011, @yahoo.com (tasked S2E, asw Iran Quds Force) has his/her Yahoo account hacked by an unknown actor, sends out spam email to his/her contact list:
Scenario: @yahoo

- @yahoo.com has a number of Yahoo groups in his/her contact list, some with many hundreds or thousands of members
- At DS-200B in particular, collection spiked as:
  - The initial spam messages were sent (and collected)
  - Inboxes of email recipients were viewed by contact list
  - Messages were sometimes viewed, but more often sent as precached views on Google and Yahoo (along with inboxes)
  - Inboxes where the recipient did not delete the spam message continued to be collected every time they were viewed
  - Some recipients added @yahoo.com to their address books (possibly as a spam defeat?) – address books were collected every time
Scenario: @yahoo

DS-200B Collection By Day - 11 Sep - 24 Sep (in MB)

DS-200B Collection By Hour – 18 Sep – 23 Sep (in MB)
Scenario: @yahoo

- **@yahoo.com** emergency detasked from DS-200B and US-3171 at 13:04Z on 20 Oct
- Numerous first-order address books and inboxes collected meant tasked selectors on address books or buddy lists of contacts of **@yahoo.com** also affected:
  - **@yahoo.com** and **@gmail.com** emergency detasked off US-3171 at 13:10Z on 20 Sep
- Memorializing to PINWALE only address books and inboxes owned by target selectors would have reduced PINWALE volumes 90%+
  - Site XKEYSCOREs would buffer data for SIGDEV purposes
  - Metadata from known owner address books and inboxes stored regardless

TOP SECRET//SI//NOFORN
Mobile IMAP

- IMAP protocol used by email clients to fetch mail from server(s)
- Not designed for devices with intermittent connections (i.e. mobile phones)
- Android implementation in particular uses a lot of bandwidth