Active Countermeasures for Email Fraud

Wentao Chen, Fuzhou Wang, Matthew Edwards
2023-07-04
Bristol Cyber Security Group
Great Investment Opportunity

My name is Rolf Michielson. I am a financial consultant/attorney. My client has a substantial amount of money $10.5M to invest in any good business that will bring good return on investment either in your country or anywhere in the world. My client is looking for a reliable, trustworthy and competent person to partner with. If you are interested, I will like you to email me the following:

1) BUSINESS PLAN
2) ANNUAL RETURN ON INVESTMENT OF THE PROPOSAL
3) BUSINESS DURATION
4) YOUR DIRECT TELEPHONE NUMBER

My client is willing to negotiate the terms and conditions to suit your best interest. [...]
## Email Fraud

<table>
<thead>
<tr>
<th>Crime Type</th>
<th>Victims</th>
<th>Losses ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced Fee</td>
<td>11,264</td>
<td>104,325,444</td>
</tr>
<tr>
<td>Lottery/Inheritance</td>
<td>5,650</td>
<td>83,602,376</td>
</tr>
<tr>
<td>Extortion</td>
<td>39,416</td>
<td>54,335,128</td>
</tr>
<tr>
<td>Investment</td>
<td>30,529</td>
<td>3,311,742,206</td>
</tr>
<tr>
<td>Confidence/Romance</td>
<td>19,021</td>
<td>735,882,192</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>105,880</td>
<td>4,289,887,346</td>
</tr>
</tbody>
</table>

\(^1\)IC3 2022 figures, see https://www.ic3.gov/
Canham & Tuthill (2022) argue:

The current state of information security can therefore be simplified as follows; high numbers of cybercrimes are committed, extremely low arrest and conviction rates present little to no deterrence to threat actors, and humans continue to be the attack vector of choice.

[...]

employing social engineering active defense (SEAD) will be more effective to countering malicious actors than maintaining the traditional passive defensive strategy.

---

Scam-Baiting

- A manual example of active countermeasures.
- Various manifestations
  - reverse-scramming/cash-baiting
  - ‘tropy-hunting’
  - ‘straight’ scam-baiting.
- Increases the false positive rate for scammers\(^3\).
- Frustrates scammers\(^4\)

---

Also:

- Volunteer/hobbyist activity.
- Takes up scam-baiter time & energy.
- Doesn’t scale.
- Not a good evidence-base.

What we want is an *automatic* scam-baiter.
Server Structure
Classifier & Template Identify fraud format; respond with randomly-selected prewritten template option.

Text Generator A GPT-Neo trained on human scam-baiting exchanges in <email,response> pairs.

Text Generator B GPT-Neo trained on mix of scam-baiting & normal email conversations.
Results
## Results

<table>
<thead>
<tr>
<th></th>
<th>Hook</th>
<th>Replies</th>
<th>Max.</th>
<th>Max. Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avg.</td>
<td></td>
<td>Max.</td>
<td></td>
</tr>
<tr>
<td>Classifier &amp; Template</td>
<td>7%</td>
<td>2.45</td>
<td>5</td>
<td>17.2 days</td>
</tr>
<tr>
<td>Text Generator A</td>
<td>5%</td>
<td>2.06</td>
<td>5</td>
<td>17.3 days</td>
</tr>
<tr>
<td>Text Generator B</td>
<td>6%</td>
<td>4</td>
<td>23</td>
<td>21.5 days</td>
</tr>
</tbody>
</table>
Concurrent Engagement

510 scam emails

374 addressable

137 unreachable

Server 1

62 responses

8 ongoing

Server 2

57 responses

11 ongoing

27 common

4 ongoing with both
### Scammer

The money is $2,500

### Scammer

The shipment is on now

### Text Generator B

Dear David, doing what I’m supposed to do, going to bring the US to Israel, maybe take some ham and gruff out of the plane if maybe it’s some hardwood muslin. But sure you will give it lot of thought. Here are my For short talk, but in the end you and are really going to have ham and wubbbut. chat.ze link. Best, Michael
### Limitations

#### Text Generator B

Welcome, uffs, goodFoo, yo boxes Best, Michael

#### Scammer

Pay the money for the shipment
How does ChatGPT do?

Very well\textsuperscript{a}. 28-day experiment testing 2 different ChatGPT prompting approaches (+ 1 control), found:

- Much improved (32-34\%) scammer response rate.
- ChatGPT established longer conversations, some lasting the entire study period.
- Over 1,000 replies from 250+ scammers.
- New conversation-stoppers.

\textsuperscript{a}(Bajaj & Edwards, in preparation).

Sample

Thank you for your email regarding the release of my funds approved by the IMF. I appreciate your prompt response and would like to know more about the processing and transfer cost of US Nineteen Thousand Five Hundred and Fifty United States Dollars Only. Can you please provide me with a breakdown of the cost and the reason for such a high amount? Also, I would like to know more about the official Bitcoin payment method that you mentioned.[...]
1-month experiment\(^5\) assigning scammers to conversations with one of 4 ‘personalities’ (+control). Total engagement: 1.5k responses, 294 scammers.

**Alex** is an optimistic, informal young man who sends bubbly messages with smiley faces and poor spelling.

**Dan** is a rude, impulsive middle-aged man who makes greedy and selfish demands.

**Sam** is a professional businessman who displays his credentials and always remains polite.

**Doris** is a kind, trusting old woman, who recently lost her husband and apologises for being slow.

*Which personality sustained **significantly** more engagement than the control?*

\(^5\)(Robinson & Edwards, in preparation)
https://github.com/scambaitermailbox/

- A corpus of historic human scam-baiter interactions (658 conversations, \(\approx\)37k messages).
- Code for running scam-baiting experiments.
- A growing corpus of automatic scam-baiting conversations. (723 conversations, \(\approx\)5k messages).
Future Directions

1. Application to media other than email.
2. Longer-term deployments.
3. Testing further hypotheses about scammer interests.
4. Designing automatic scam-baiters to gather information.
5. Methods robust to detection techniques.
Thank you

Questions?