The role of phone numbers in understanding cyber-crime

A. Costin * J. Isachenkova * M. Balduzzi +
A. Francillon * D. Balzarotti *

*Eurecom, Sophia Antipolis, France
+Trend Micro Research, EMEA

July 11, 2013
Introduction

Online/digital identifiers in cyber-crime

- Mail
- Domain name/Web site
- Social networks/Nicknames profile
  - Extensive studies: \([\text{LMK}^{+10}, \text{TGM}^{+11}, \text{KKL}^{+08}, \text{Ede03}, \text{CHMS06}]\)
- Phone numbers
  - Limited studies: \([\text{CYK10}, \text{STHB99}, \text{Pol05}, \text{Hyp}]\)
  - Studied mainly in context of premium short number mobile frauds
  - Our main focus
Phone number usages

- Mail signatures
- Extensively used in many businesses
- Offers less anonymization than other identifiers
- Links cyber domain to reality domain
- Commonly used in various online frauds, e.g.:
  - Premium numbers fraud
  - Scam fraud
Introduction

Importance of cyber-crime and phone numbers – example

- Banking Trojan. Shylock [symb]
- Injects code into banking websites
- Replaces telephone details into the contact pages of online banking websites

<table>
<thead>
<tr>
<th>Injected Numbers</th>
<th>Original Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calling from the UK</td>
<td>Calling from abroad</td>
</tr>
<tr>
<td>0800 310</td>
<td>+44 8705</td>
</tr>
<tr>
<td>0800 310</td>
<td>+44 118 9</td>
</tr>
<tr>
<td>0800 310</td>
<td>+44 8705</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Calling from the UK</th>
<th>Calling from abroad</th>
</tr>
</thead>
<tbody>
<tr>
<td>08457</td>
<td>+44 8705</td>
</tr>
<tr>
<td>0845 3</td>
<td>+44 118 9</td>
</tr>
<tr>
<td>08457</td>
<td>+44 8705</td>
</tr>
</tbody>
</table>
Hypothesis

- Phone numbers are used in cyber-crime activities
- Can we find telecom operators preference?
- Can we find geographical preference?
- Phone numbers can be a stronger identification metric vs. other identifiers
Goals

- Check those hypothesis against real data-sets
- Evaluate the reliability of automated phone numbers extraction and analysis
  - Identify challenges and limitations
- Automatically find patterns associated with recurrent criminal activities
- Automatically correlate the extracted information for
  - Telecom operator preference
  - Geographic area preference
Methodology

Sources

- SCAM
- SPAM
- DNS
- ANDROID

Data processing

- parse phone detect
- phone extract
- language heuristic
- keywords heuristic
- patterns heuristic
- phone normalize
- confidence level

Data analysis

- graphs
- business models
- communities
- HLR analysis
- google PRank

DB storage queries
Datasets I

Data Sources initially considered

- SPAM
  - Large and extremely noisy dataset
  - Extremely challenging to extract and clean phone numbers

- WHOIS
  - Focused on malicious domains
  - High quality dataset (intl. format)
  - Phone numbers are dummy or replaced by CERTs’ contact numbers
Datasets II

- **ANDROID**
  - Small and noisy dataset
  - Mainly contained *short premium numbers* – open problem

- **SCAM**
  - Large and high quality dataset
  - Phone numbers are an important part of business model
  - Focus on this dataset
Phone Number Extraction I

Success and Reliability of Extraction depend on

- How well formatted the number is
  - Call: 0336 9505705 9 am - 5 pm
  - Can be decoded as 2 valid numbers: +443369505705 or +33695057059
  - We aim at obtaining:
    - Non ambiguous normalized number
    - Fully qualified international format number
Phone Number Extraction II

- How structured and easy to parse the information is
  - WHOIS records (easy) vs.
  - Malicious mobile binary (difficult)
- How noisy the data source is
  - Spam messages are very noisy (to defeat anti-spam filters)
  - Scam messages have almost no noise
Phone Number Extraction Challenges

Example Number obfuscation used [syma]

\[(Ч^95)1 ^2^ три ^40^ ОО (495) один 2 3 − 4 0 − 00 \} \quad (495)123400\]

\[/495/ Ч;3 =2 l 00 0 (Ч~9~5) 43~2~ один~0~0~0 \} \quad (495)432100\]
Subject: BUSINESS PROPOSAL
Date: Tue, 26 Jan 2001 22:13:10 -0700 (PDT)
From: emmanuel udo <emmanuel2@yahoo.com>
To: tm.richardson@arcanum.ca

FROM: DR. EMANUEL UDO.
TEL: 234 1 759 1549; FAX: 234 1 759 0179.
E-MAIL: emmanuel2@yahoo.com

BUSINESS PROPOSAL

ATTN: PRESIDENT / CEO,

My name is EMANUEL UDO, a member of the Presidential Task Force on Oil Spillage Clean-up.
Early last year, there was a major oil spillage in the Niger Delta Region of Nigeria which rendered over 70% of the communities homeless. The contract was handled by a foreign firm but because of the huge monetary profit we envisaged, we decided to over-invoice the contract sum. Now the contract has been completed and the original contractor has since been paid, but the contract balance of US$88 million, which resulted from the over-invoiced contract sum that has been left in a suspense account with the CENTRAL BANK OF NIGERIA, is what we and my partners are planning to take out of
SCAM Dataset

- Used *user reports aggregator* 419scam
- Data timespan: January 2009 – August 2012
- Enriched and correlated with numbering plans (NNPC) databases
  - Free (*libphonenumber*)
  - Commercial (more detailed and updated)
SCAM Email Categories

- Emails classified in 10 categories
- 3 categories cover over 90% of the data

Scam Categories

- Financial scam (62%)
- Fake lottery (25%)
- Next of kin (8%)
- Other (5%)
SCAM Phones Categories

- ~67k unique normalized phone numbers
- Classified using numbering plans (NNPC) databases

Number type breakdown

- UK PRS (51%)
- Mobile (44%)
- Other (5%)
SCAM Communities/Identity Links

- Used clustering techniques, discovered identity links
- Identified 102 communities
- Supports the hypothesis that phone numbers are a good metric to study scammers
ANALYSIS OF MOBILE PHONE NUMBERS
Questions and Hypothesis

- For how long are phone numbers used?
- Are phone numbers reused or discarded?
- If discarded, after how long?
- Are phone numbers used in roaming?
- If roaming, to which extent?

We try to answer these questions with HLR queries.
HLR Querying

- HLR = Home Location Register
- Important component of Mobile Network Operators

![HLR Query Diagram](Image)
In Aug 2012, querying once for all mobiles encountered in:
- Jan – Jun 2012
- Jul 2012
Repeated HLR Queries

Performed HLR queries

- For 1400 numbers
- Every 3 days
- During Jul – Aug 2012
- Hypothesis 1: Possibility of a link with the Nigerian groups
- Hypothesis 2: May be used to conceal location
Question:

- For how long a scam number is used?

Phone number reuse

Age of reused phone number (years)

Reuse percentage

- 3
- 2
- 1

Reuse percentage

- 18
- 16
- 14
- 12
- 10
- 8
- 6
- 4
- 2
- 0

Age of reused phone number (years)
ANALYSIS OF UK PRS PHONE NUMBERS
Definition

- Premium rate services (PRS) are a form of micro-payment for paid content, data services and value added services that are subsequently charged to user phone bill
- UK PRS is a 800 Mil. GBP business (2009)
What are UK PRS numbers? II

Usages

- Conceal geographic location of real phone, via *call forwarding*
- Earn revenue from calls to these numbers

Challenges

- Hard to trace the ”service provider”
- Hard to trace the real phone number behind forwarding
- Hard to detect or prove that fraud is involved
Range of UK PRS numbers

- ~34k unique phone numbers in UK range of 07x
- Premium Rate Services numbers
- 4 operators (out of 88) provide more than 90% of fraud-related UK PRS numbers
- ~5% of one operator allocated range is fraud-related
Conclusion – Results

- Phone numbers are a strong digital identifier in some cyber-crime activities
- Phone numbers help in automated scammer community detection
- HLR lookups help
  - in identifying recurrent cyber-criminal business models
  - to study phone numbers’ geographical use and activity patterns
Conclusion – Future Work

- Phone number extraction is an open, non-trivial problem
  - Improve matching algorithms and their context-awareness
- PRS phone numbers are opaque
  - Is a "traceroute" of PRS phone numbers possible?
  - Learn business models behind them
- Short number extraction and evaluation
  - Open and challenging, non-trivial problem
  - Becomes a growing concern with mobile malware
Questions?

Contacts:

- Software and System Security Group @ EURECOM
- S3.eurecom.fr

Thank you!
References I


References II

Mikko Hypponen, *Malware Goes Mobile*,


References III


Trojan. Shylock Injects Phone Numbers into Online Banking Websites,