Using Cross-Media Relations to Identify Important Communication Requests:
Testing the Concept and Implementation

Kumiko Ono and Henning Schulzrinne
{kumiko, hgs}@cs.columbia.edu
Outline

1. Introduction
   • Motivation
   • Challenge and approaches
   • Hypothesis
   • Proposed mechanisms

2. Implementation as proof of concept

3. Observing email: testing the concept
Motivation: Receiving unwanted calls

- Important calls with an unknown caller ID, mistakenly labeled “unwanted”
  - Originating from persons/organizations connected with weak social ties
Challenge and approaches

- **Challenge:** How to identify unwanted and non-spam calls from calls shown in gray

- **Approaches**
  - Enhance white listing
  - Focus on *prior contact* through different communication means
  - “Cross-media relations”
  - e.g., email messages prior to making a call
Hypothesis

• A significant fraction of incoming calls are non-spam with an unknown caller ID.
  - From persons/organizations connected with weak social ties
    • Usually not in callee’s address book
  • Difference between a spammer and a legitimate caller
    - A spammer makes a call with no prior contact with the callee.
    - A legitimate caller has prior contact before making a call except in emergency cases.
      • A legitimate caller often transitions:
        • Web transactions → email /instant messaging → voice calls
        • Web transactions → voice calls
Hypothesis

- A significant fraction of incoming calls are non-spam with an unknown caller ID.
  - From persons/organizations connected with weak social ties
    - Usually not in callee’s address book
- Difference between a spammer and a legitimate caller
  - Prior contact via web/email/others (cross-media relations) is a distinguishing feature between a spammer and a non-spammer, except in emergency cases.
  - A legitimate caller often transitions:
    - Web transactions → email/instant messaging → voice calls
    - Web transactions → voice calls

Kumiko Ono@IPTComm
Proposed mechanisms: Using cross-media relations

- Two mechanisms based on how the callee uses prior contact
  1. Collecting as many contact addresses of potential callers as possible
  2. Providing potential callers with a weak secret as a proof of prior contact

<table>
<thead>
<tr>
<th>1. Contact addresses: Information provided by potential callers</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Web-then-call: Contact addresses in plain text or hash format</td>
</tr>
<tr>
<td>b. Email-then-call: Contact addresses</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Weak secret: Information provided by callee</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Web-then-call: Customized contact address of the callee</td>
</tr>
<tr>
<td>b. Email-then-call: Message-ID of an outgoing email message</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
   • Motivation
   • Challenge and approaches
   • Hypothesis
   • Proposed mechanisms

2. Implementation as proof of concept

3. Observing email: testing the concept
CURE system

1. Contact addresses of potential callers
2. A weak secret as a proof of prior contact

Kumiko Ono@IPTComm
CURE system

Firefox Addons

Apache supporting HTTP-EQUIV tag

Web Server

Potential callers
e.g., airline

Email

Alice

Web Browser

Email

User of CURE System
Bob

IMAP clients

DBMS: MySQL
API: REST, JSON

OpenSER

SIP communicator supporting “Sender-Ref” header in INVITE

Inbound SIP Server

Contact addresses of potential callers

1. Cross-media relations data

2. A weak secret as a proof of prior contact

Prior contact
2-a. Using a weak secret: Web-then-call

Use **a random component** since no transaction ID in HTTP

W1) When sending a sign-up form:
   HTTPS POST request
   phone=$\text{sip:user+SDJP09lk@columbia.edu}$

the same as email subaddressing

C1) INVITE
   From: Anonymous
   To: $\text{user+SDJP09lk@columbia.edu}$

W2) Update

C2) Query

C3) Accept or Decline

User of CURE System

Bob

Cross-Media Relations data

Web Browser

Add-on

Web Server

e.g., airlines

Potential Callers

Alice

Inbound SIP Server

Kumiko Ono@IPTComm
2-b. Using a weak secret: Email-then-call

Use the Message-ID of outgoing message

E1) Sending a message
Message-ID: 004301c9b17f257f6a40707e3ec0@columbia.edu

C1) INVITE
From: Anonymous
Sender-Ref: 004301c9b17f257f6a40707e3ec0@columbia.edu

C2) Query
Inbound SIP Server

C3) Accept or Decline

User of CURE System
Bob

Mail UA

MDA

Cross-media relations data

E2) Update

Mail protocols e.g., SMTP/IMAP

SIP

DB API
Demo: 2-a. Using a weak secret in web-then-call

W1) When sending a sign-up form:
HTTPS POST request
phone=sip:user+SDJP09lk@columbia.edu

the same as email subaddressing

C1) INVITE
From: Anonymous
To: user+SDJP09lk@columbia.edu

C2) Query
Inbound SIP Server

C3) Accept or Decline
Cross-Media Relations data

W2) Update
Web Browser Add-on

User of CURE System
Bob

Potential Callers

Web Server
e.g., airline

Alice
Kumiko Ono@IPTComm 11
Demo: 2-a. Using a weak secret in web-then-call

Screenshot of Firefox Add-on connecting to opentable.com

Your SIP phone: Enter email

Your Email: *bob+anwxy3fa@cs.columbia.edu

Record #bob_cure@cs.columbia.edu

- user: bob_cure@cs.columbia.edu
- doAccept: 1
- description: secure.opentable.com/register.aspx
- url: secure.opentable.com/register.aspx
- contactType: sip
- contact: bob@cs.columbia.edu
- delimiter: +
- secret: anwxy3fa
Demo: 2-a. Using a weak secret in web-then-call

Record #bob_cure@cs.columbia.edu

user bob_cure@cs.columbia.edu
doAccept 1
description secure.opentable.com/register.aspx
url secure.opentable.com/register.aspx
contactType sip
contact bob@cs.columbia.edu
delimiter +
secret anwvx3fa
timestamp 2010-11-30 13:47:58
Outline

1. Introduction
   • Motivation
   • Challenge and approaches
   • Hypothesis
   • Proposed mechanisms

2. Implementation as proof of concept

3. Observing email: testing the concept
Testing the concept

- Ideally, evaluate the concept using the implementation
  - But...
    - Low volume of unwanted calls
    - Need cooperation of web sites
    - Need end-to-end SIP connections
- Instead, observing incoming email messages
  - Stored email messages easier to categorize than call history or CDRs.
Survey of incoming email

- Participants: our colleagues and other students in CU
- Data set: their email messages as substitutes for CDRs
  - Headers of incoming messages for 4 weeks in March 2010
  - Collected by providing a dedicated IMAP client for this survey
  - 7575 messages received and stored by 12 email accounts
    - 3618 messages for 5 university email accounts
    - 3967 messages for 7 free email accounts
- Methodology:
  - Categorize messages into groups
- Metric: fraction of incoming messages in each group
Categorizing incoming messages

- An incoming message
  - Sender addr. found in outgoing msgs.?
    - Yes
      - Sent earlier
      - Replied within two weeks
    - No
      - Sender addr. found in user's config.?
        - Yes
          - Trusted users
        - No
          - Sender addr. found in social graphs based on msgs.?
            - Yes
              - FoF found in CCed msgs.
            - No
              - FoF found in msgs. from mailing lists
    - No
      - Have prior contact with the sender?*
        - Yes
          - Web
          - Email
          - Call
        - No
          - Famous?*
            - Yes
              - Publicity
            - No
              - Unwanted msg.?*
                - Yes
                  - Unwanted (spam)
                - No
                  - Don't know
                    - Unlabeled

*Manually determined by participant

Kumiko Ono@IPTComm 11
Fractions of messages in groups

- Web-then-email appears very effective. (52% on average)
  → Web-then-call would also be.
  → Email-then-call?

- Varies according to the account usage

Using cross-media relations appears to be effective as another tool for identifying non-spam communication requests.
Summary

• Using cross-media relations to identify non-spam communication requests
  - Survey shows 52% of incoming email have unknown sender addresses but having web-then-email relations
    - Useful as additional component of call filtering system

• To provide more evidence of effectiveness
  - Survey of received email messages/calls/SMSes
    - *-then-email, *-then-call, *-then-SMS
    - Take part in survey at https://irt-win7.cs.columbia.edu/