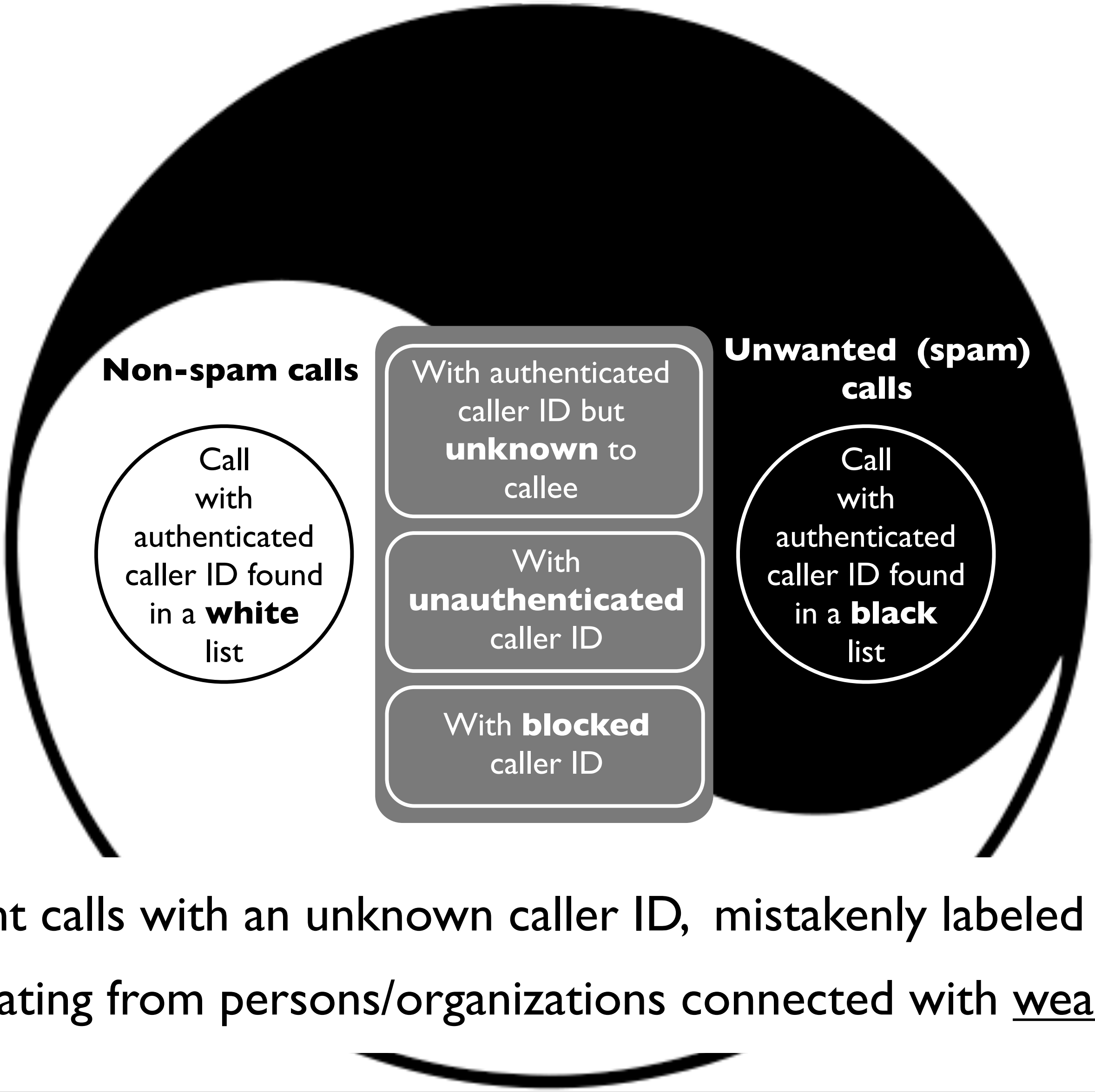
Using Cross-Media Relations to Identify Important Communication Requests: Testing the Concept and Implementation

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

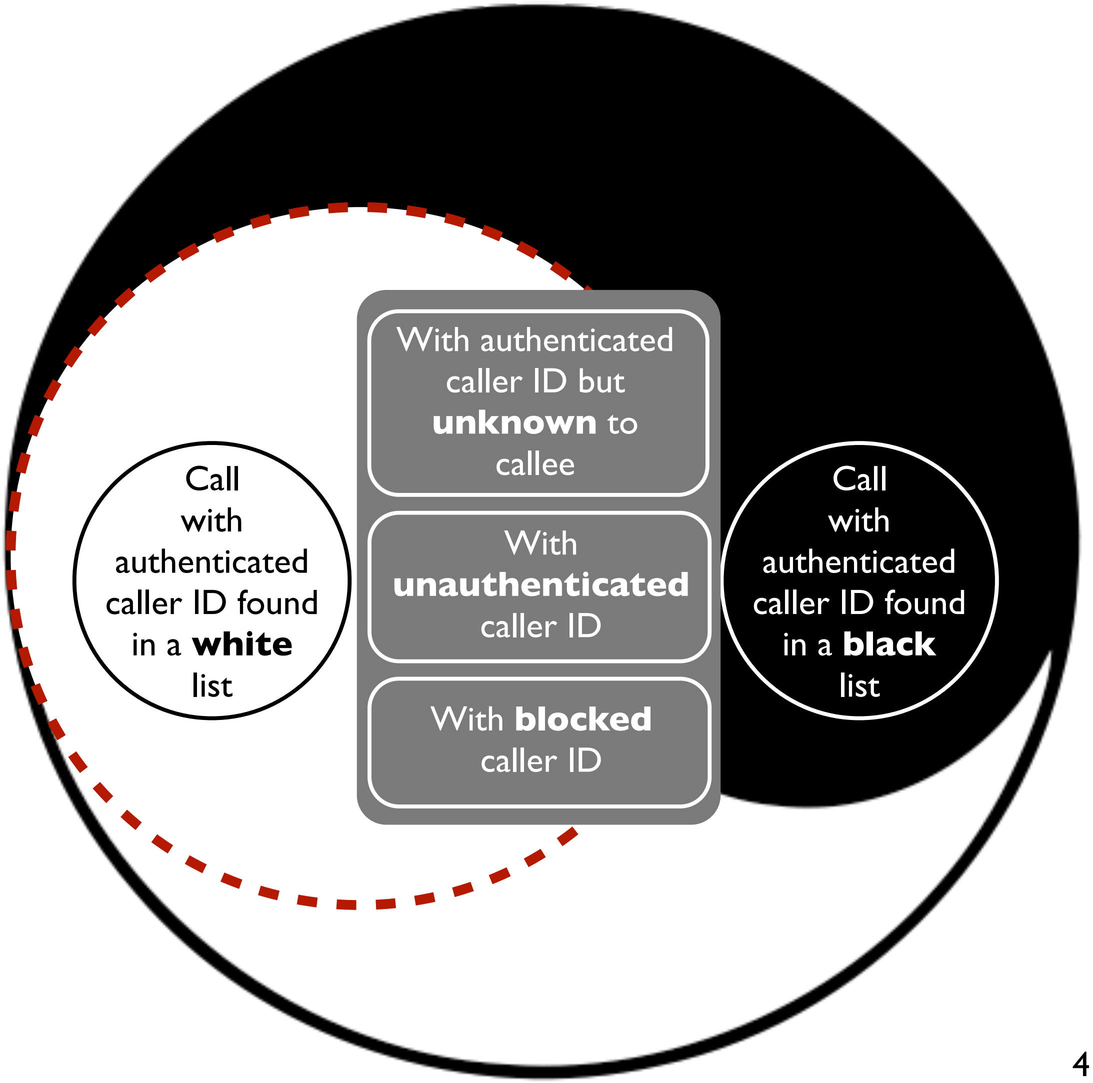
Motivation:
Receiving unwanted calls


Non-spam calls
Call with authenticated caller ID found in a **white** list
With authenticated caller ID but **unknown** to callee
With **unauthenticated** caller ID
With **blocked** caller ID
Unwanted (spam) calls
Call with authenticated caller ID found in a **black** list

- Important calls with an unknown caller ID, mistakenly labeled "unwanted"
  - Originating from persons/organizations connected with <u>weak social ties</u>

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


Call with authenticated caller ID found in a **white** list
With authenticated caller ID but **unknown** to callee
With **unauthenticated** caller ID
With **blocked** caller ID
Call with authenticated caller ID found in a **black** list

Hypothesis

- A significant fraction of incoming calls are non-spam with an unknown caller ID.
  - From persons/organizations connected with <u>weak social ties</u>
    - 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.

  ee.

  except in emergency cases.
    - A legitimate caller often transitions:
      - Web transactions → email /instant messaging → voice calls
      - Web transactions → voice calls

Proposed mechanisms: Using cross-media relations

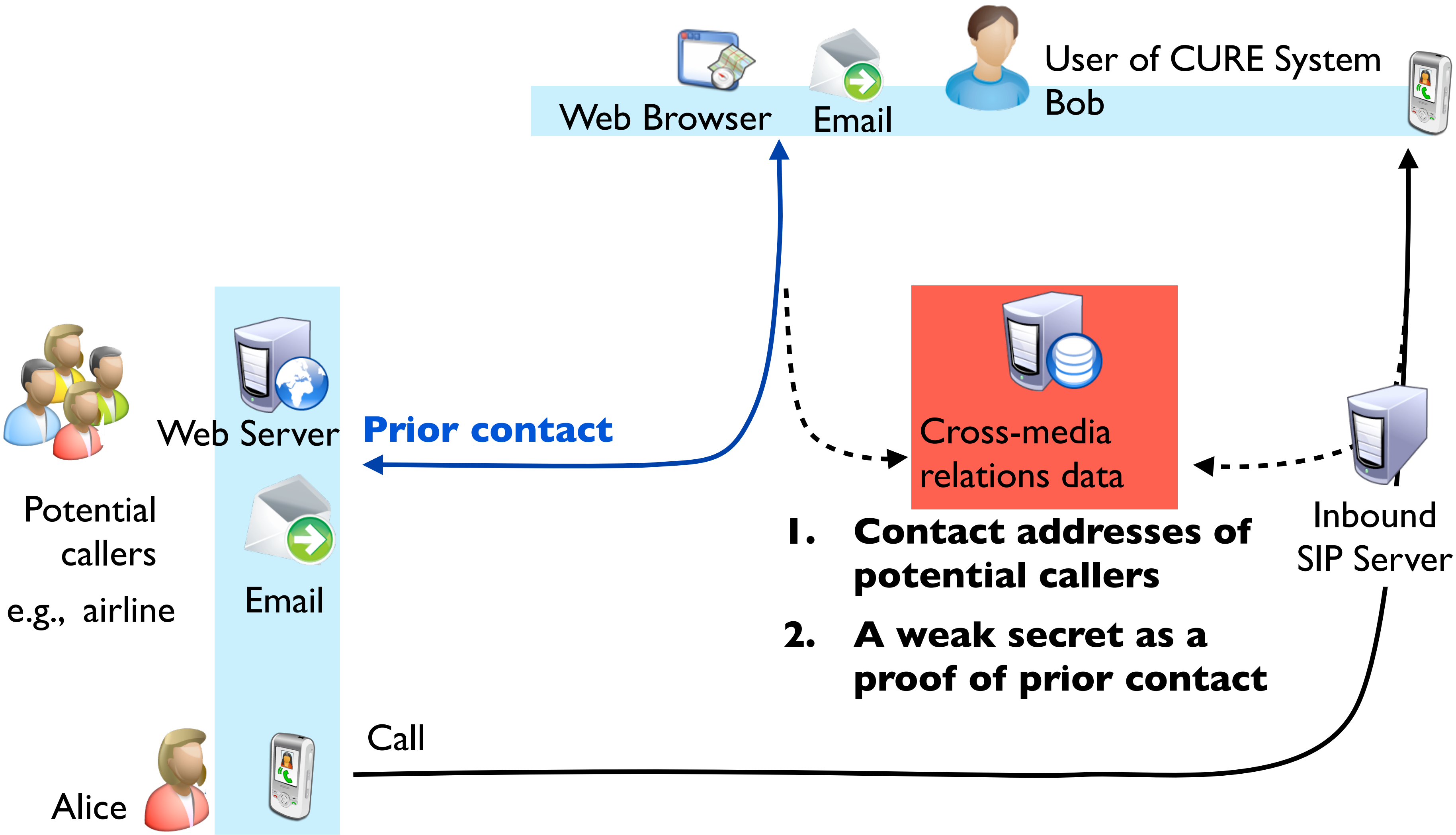
- Two mechanisms based on how the callee uses prior contact
  1. Collecting as many **contact addresses of potential callers** as possible
  2. Providing potential callers with a **weak secret** as a proof of prior contact

1. Contact addresses: Information provided by potential callers	**2. Weak secret:** Information provided by callee
a. Web-then-call: Contact addresses in plain text or hash format b. Email-then-call: Contact addresses	a. Web-then-call: Customized contact address of the callee b. Email-then-call: Message-ID of an outgoing email message

Outline

1. Introduction
   - Motivation
   - Challenge and approaches
   - Hypothesis
   - Proposed mechanisms
2. Implementation as proof of concept
3. Observing email: testing the concept

CURE system



Web Browser

Email

User of CURE System Bob

Potential callers

e.g., airline

Web Server

Email

Alice

Prior contact

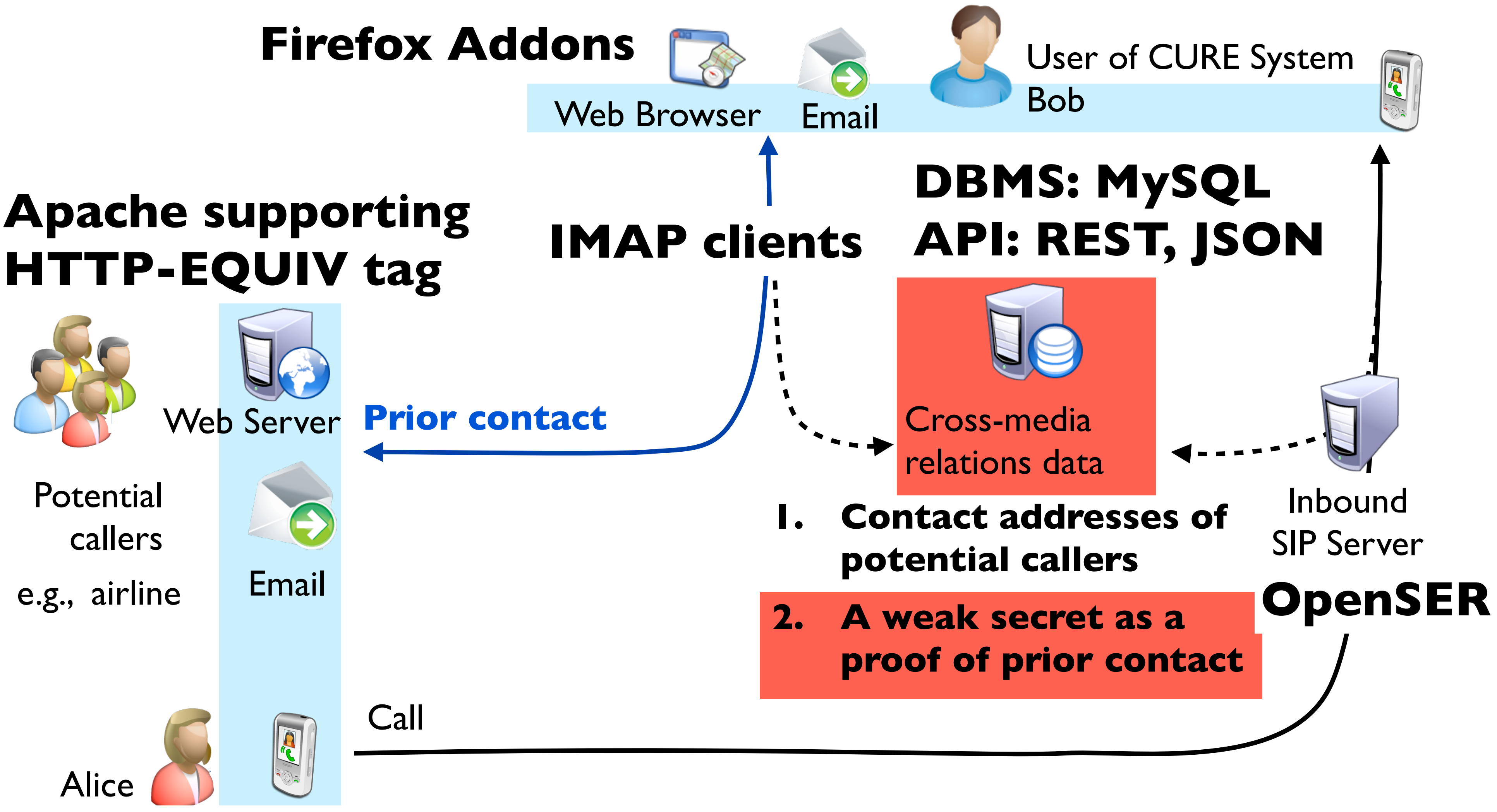
Cross-media relations data

Inbound SIP Server

Call

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

CURE system



Firefox Addons

Web Browser

Email

User of CURE System
Bob

Apache supporting HTTP-EQUIV tag

IMAP clients

DBMS: MySQL
API: REST, JSON

Web Server

Prior contact

Cross-media relations data

Potential callers

e.g., airline

Email

Inbound SIP Server

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

OpenSER

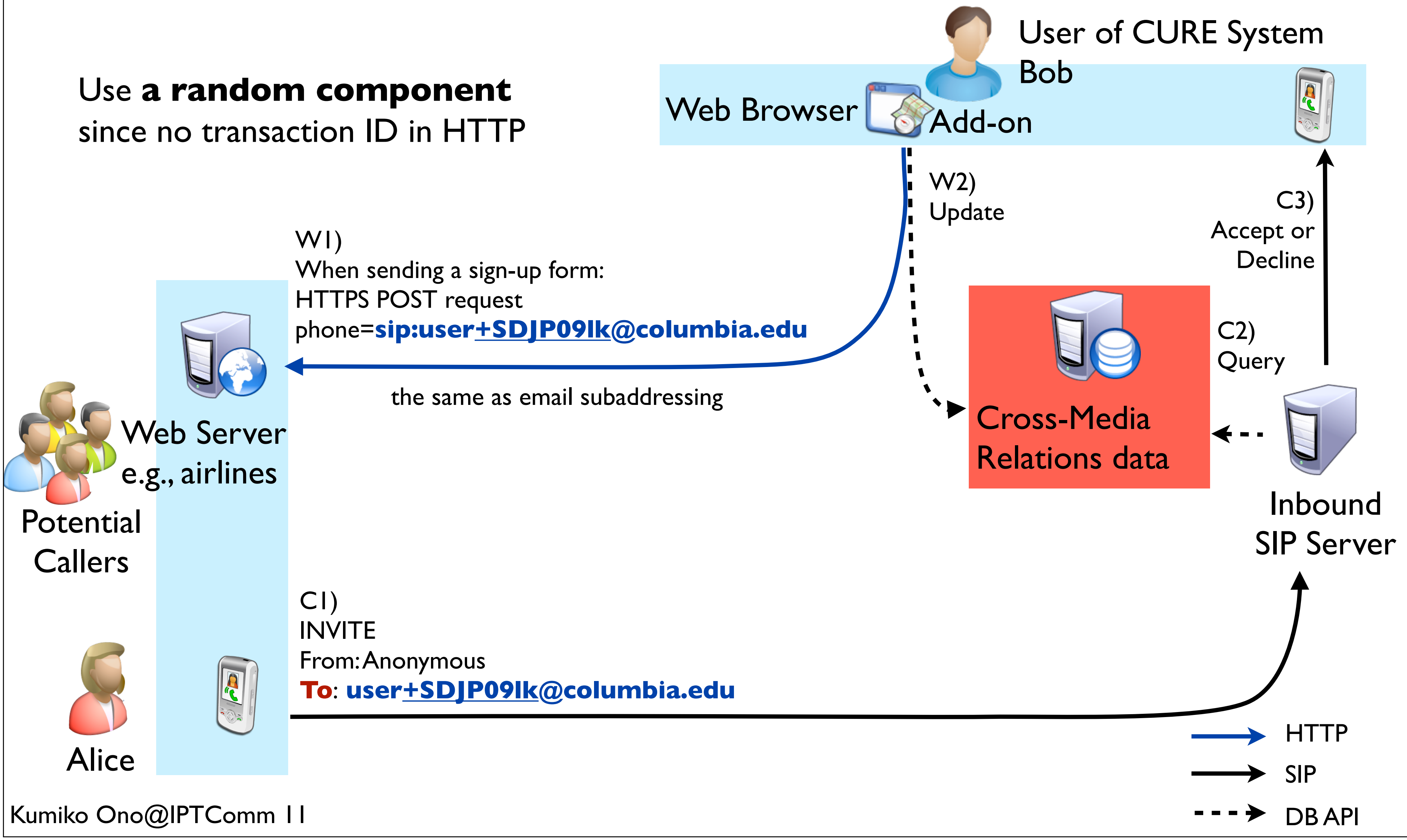
Call

Alice

SIP communicator supporting “Sender-Ref” header in INVITE

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

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



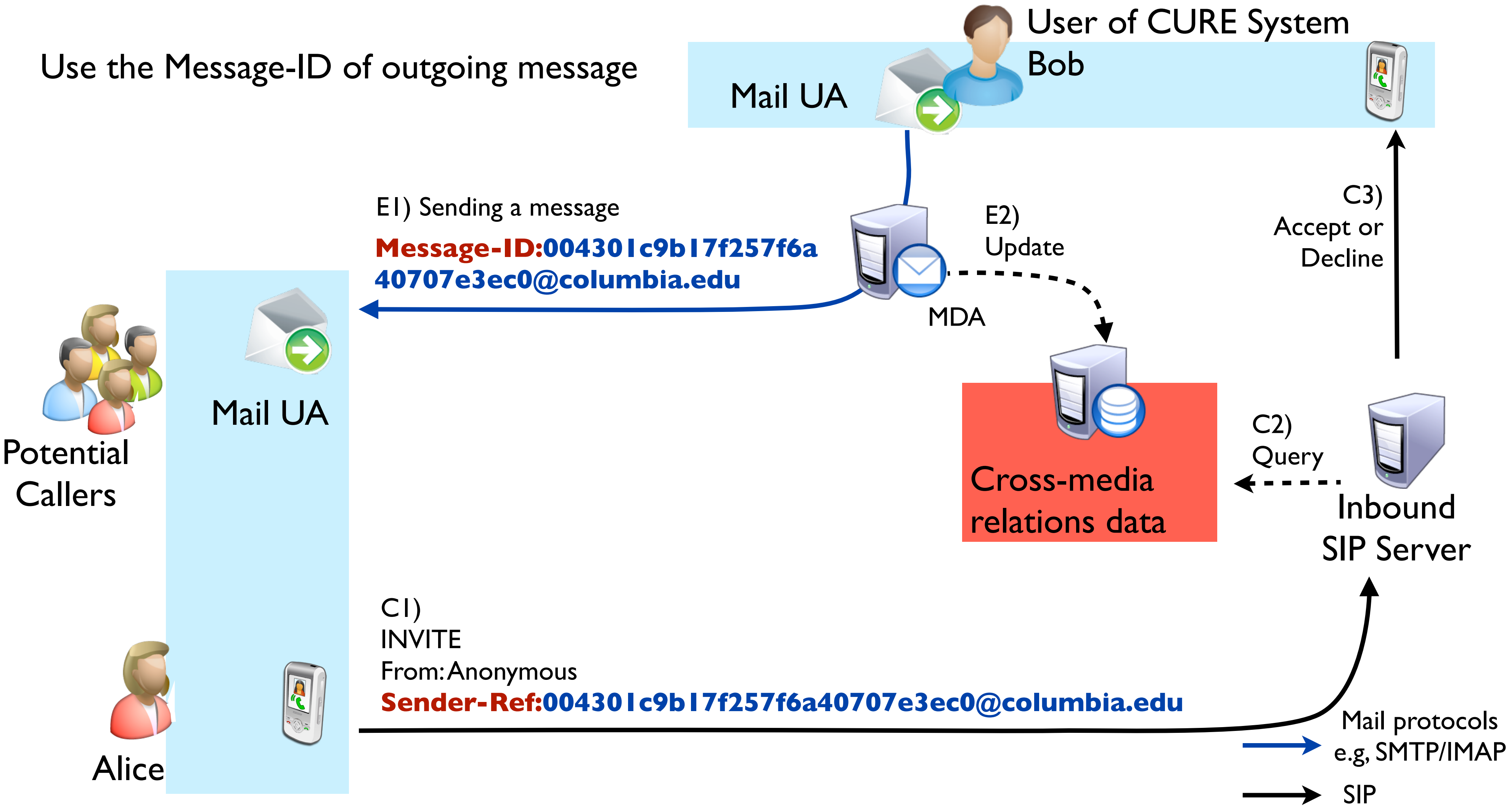
W1)
When sending a sign-up form:
HTTPS POST request
phone=**sip:user+SDJP09lk@columbia.edu**

the same as email subaddressing

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

2-b. Using a weak secret: Email-then-call

Use the Message-ID of outgoing message



User of CURE System

Bob

Mail UA

E1) Sending a message

Message-ID:004301c9b17f257f6a 40707e3ec0@columbia.edu

E2) Update

MDA

C3) Accept or Decline

Potential Callers

Mail UA

Cross-media relations data

C2) Query

Inbound SIP Server

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

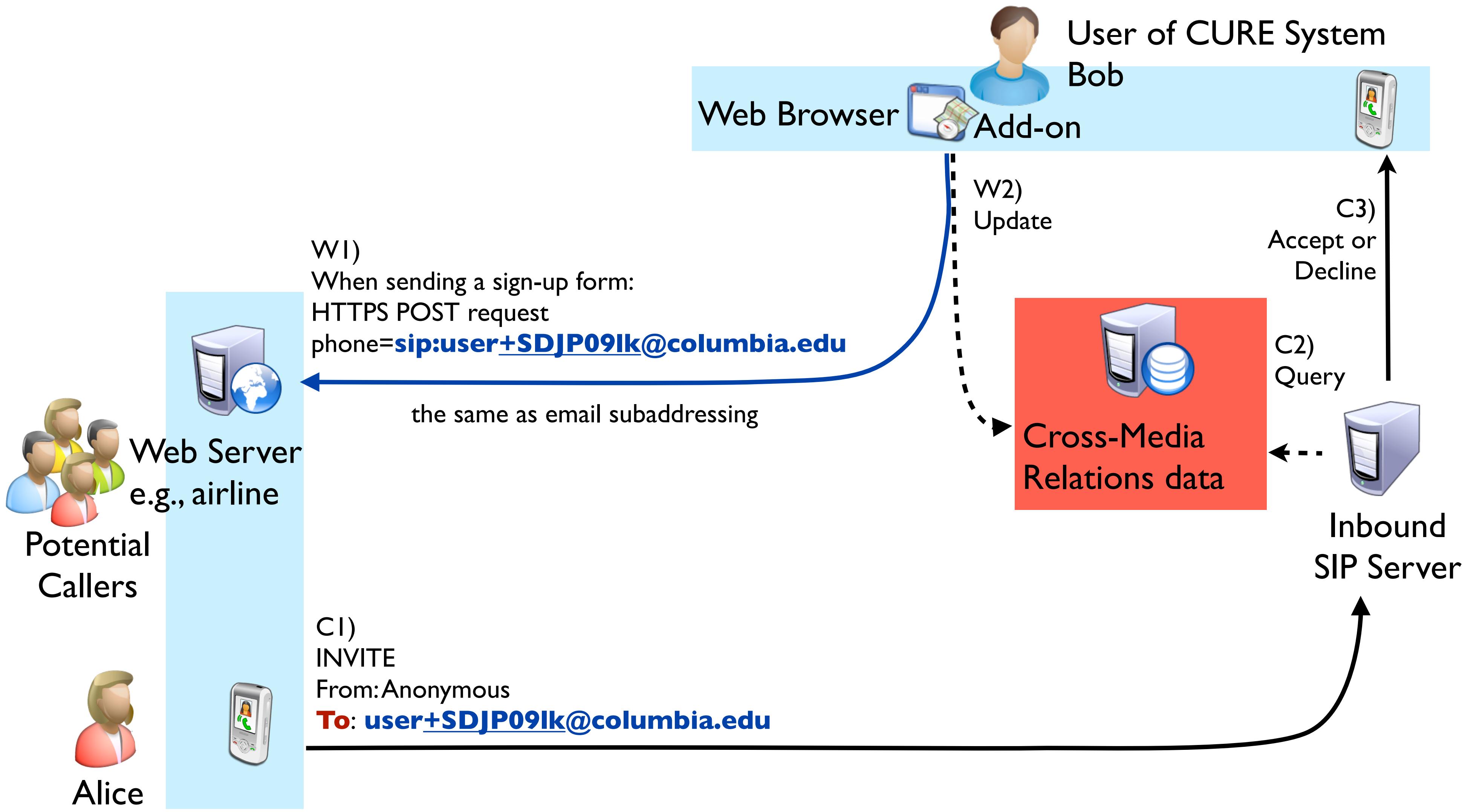
Alice

Mail protocols e.g, SMTP/IMAP

SIP

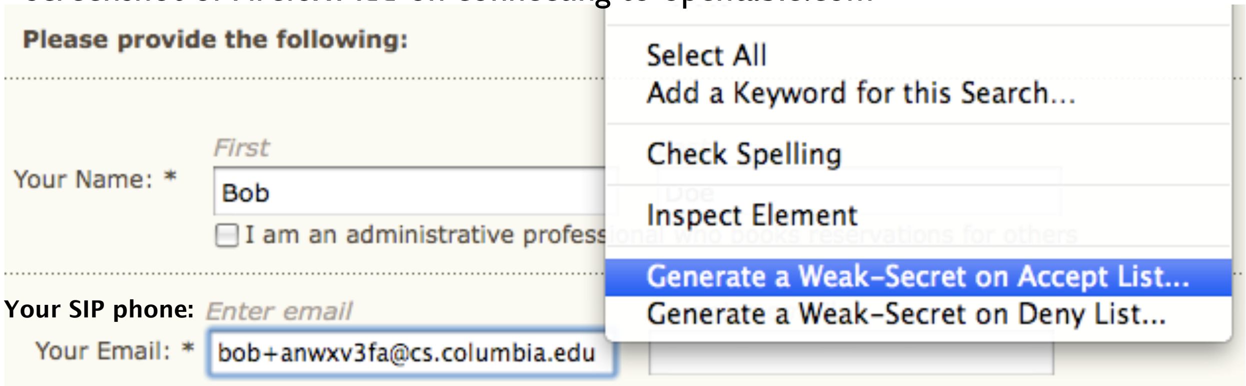
DB API

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



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

Screenshot of Firefox Add-on connecting to opentable.com



Please provide the following:

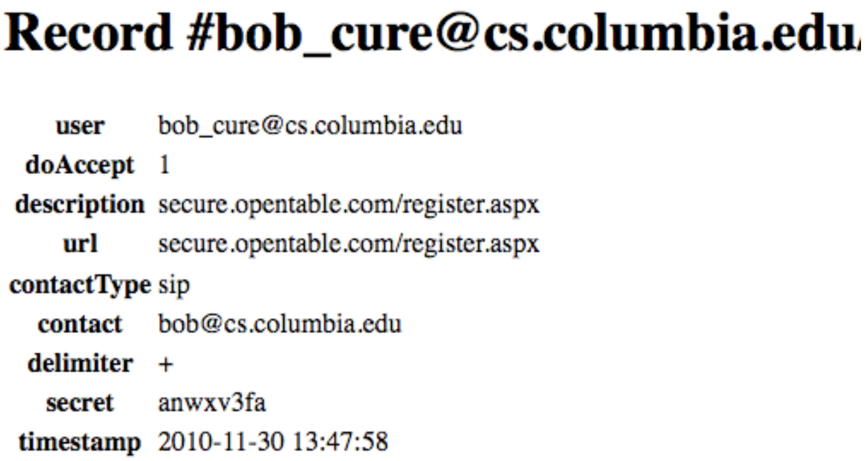
Your Name: * First: Bob

□ I am an administrative professional who books reservations for others

Your SIP phone: Enter email

Your Email: * bob+anwxv3fa@cs.columbia.edu

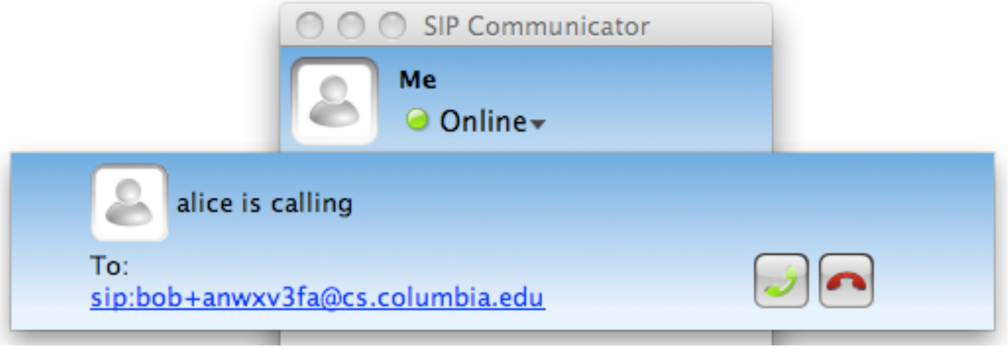
Select All
Add a Keyword for this Search…
Check Spelling
Inspect Element
Generate a Weak-Secret on Accept List…
Generate a Weak-Secret on Deny List…



Record #bob_cure@cs.columbia.edu

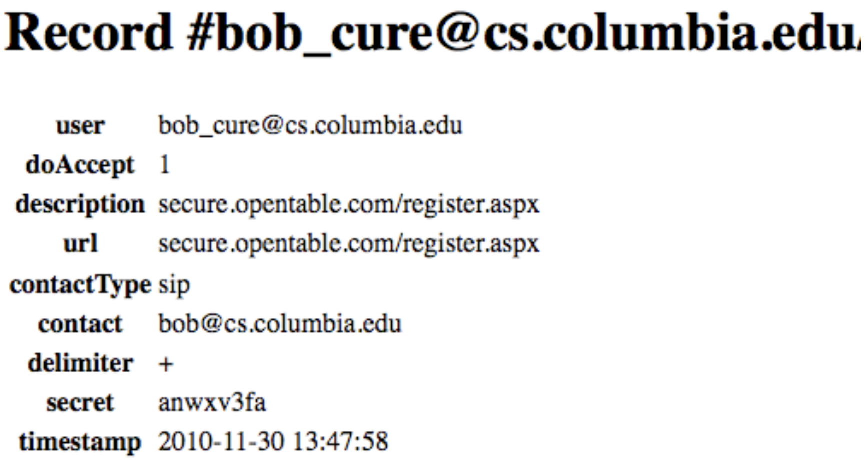
user	bob_cure@cs.columbia.edu
doAccept	1
description	secure.opentable.com/register.aspx
url	secure.opentable.com/register.aspx
contactType	sip
contact	bob@cs.columbia.edu
delimiter	+
secret	anwxv3fa
timestamp	2010-11-30 13:47:58

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



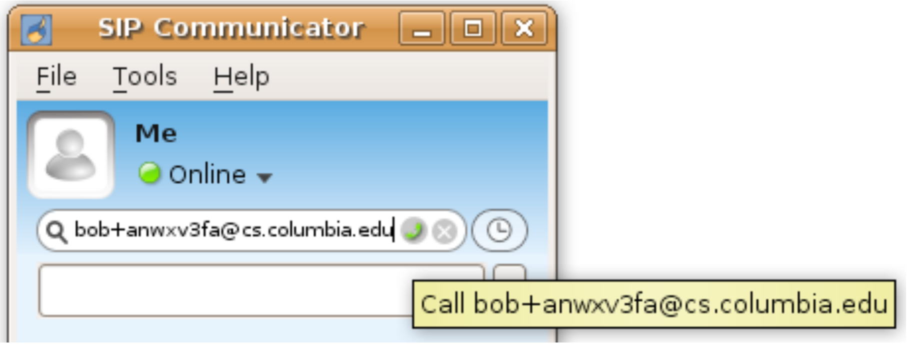
alice is calling

To:
sip:bob+anwxv3fa@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	anwxv3fa
timestamp	2010-11-30 13:47:58



bob+anwxv3fa@cs.columbia.edu

Call bob+anwxv3fa@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

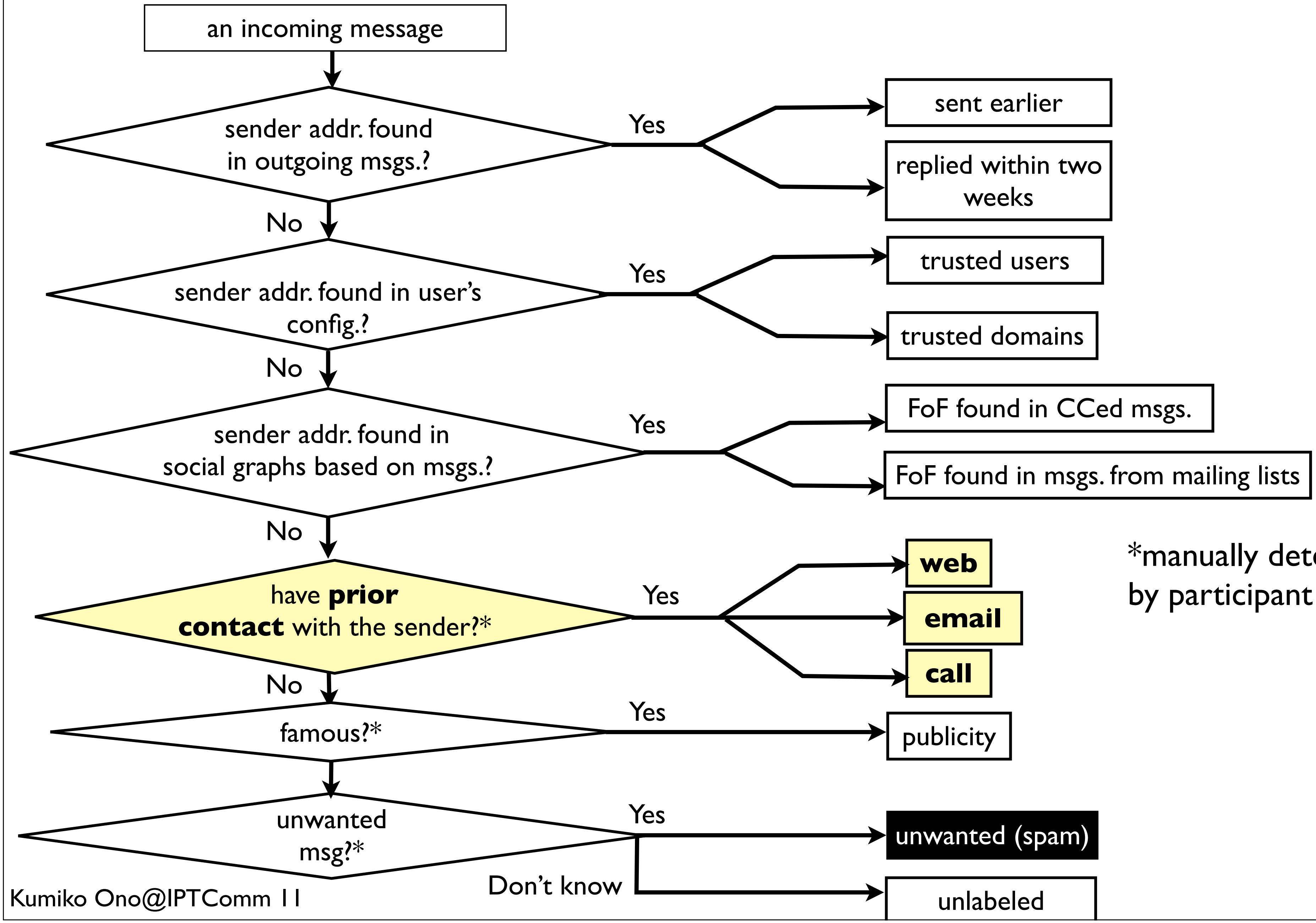
Testing the concept

- 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

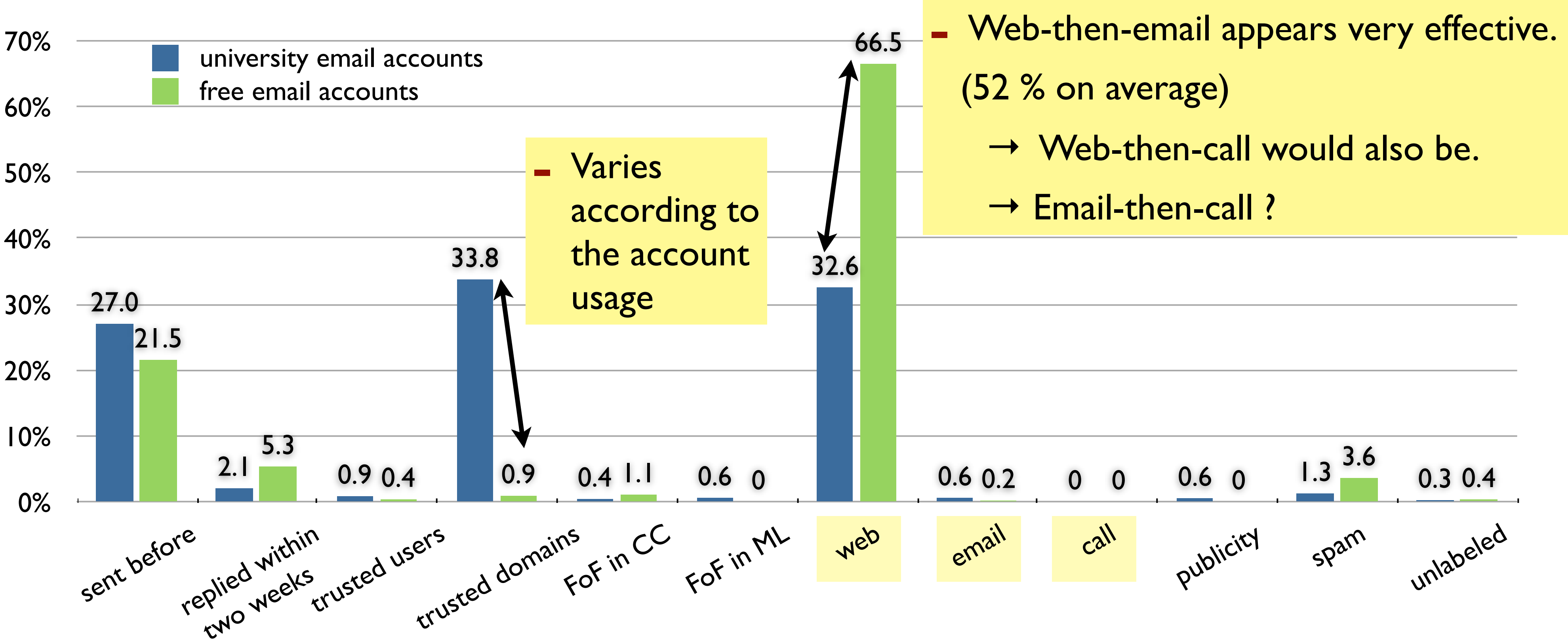
- Participants: our colleagues and other students in CU
- Data set: their email messages as substitutes for CDRs
  - Headers of incoming messages for 4 weeks in March 2010
  - Collected by providing a dedicated IMAP client for this survey
  - 7575 messages received and stored by 12 email accounts
    - 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



*manually determined by participant

Fractions of messages in groups



- Varies according to the account usage

- Web-then-email appears very effective.

  (52 % on average)

  → Web-then-call would also be.

  → Email-then-call ?

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/