Author Identification for Malicious Threat Detection in Social Media

**Presenter:**
Cihan Tunc

**Lead faculty:**
Dr. Salim Hariri, UA

**Students:**
Sicong Shao

**CAC faculty:**
Dr. Cihan Tunc
Dr. Salim Hariri

Cloud and Autonomic Computing Center
Semi Annual IAB Meeting, April 23-25, 2017
Tucson, Arizona
Outline

- Project Goals, Motivations
- Project Overview
- Project team members
- Background and Related Research
- Overview of project tasks
- Activities and outcomes
- Deliverables and benefits
- LIFE form input
Project Team Members

Faculty

Dr. Salim Hariri  
Director, CAC & ECE Department, The University of Arizona

Dr. Cihan Tunc  
Research Assistant Professor, ECE Department, The University of Arizona

Students

Sicong Shao  
PhD student, The University of Arizona
Motivation

- The user with illegitimate purposes can exploit the powerful dissemination feature of social network platform to spread malicious information and influence people opinion.
- With the anonymous nature, Internet Relay Chat (IRC) is widely used for cybercrime including hacking, cracking, and carding.
- Performing author identification on IRC is more challenging task, due to difficulties of real-time data collection, synchronous computer-mediated communication, many sophisticated suspects that are needed to be considered.
Project Goals, Motivation

Goal

- Creating autonomic IRC monitoring bot for continuous monitoring and performing pre-processing for unstructured data in real-time.
- Developing an effective personality feature extraction unit.
- Designing a robust learning unit for creating author identification model.
- Performing author identification to identify anonymity.
Project Goals, Motivation

Mastermind behind sophisticated, massive botnet outs himself

Andromeda kingpin is identified by his ICQ number.

DAN GOODIN - 12/5/2017, 2:40 PM

Malware as a service

Andromeda was primarily a service provided to other online criminals that made it easy for them to quickly spread their malicious wares. It allowed customers to build custom plug-ins for keylogging and rootkits for as little as $150, or it could serve as a platform for installing existing malware, including the Petya and Cerber ransomwares; the Neutrino bot for DDoS attacks; information-stealing malware known as Ursnif, Carberp, and Fareit; and the Lethic spam bot. The botnet network relied on more than 1,200 domains and IP addresses to control infected computers. Over the past six months, Microsoft detected or blocked the Andromeda bot on more than one million computers every month on average.
Previous Work

- There are various studies in identifying the author information
  - For plagiarism checking
  - For understanding if the accounts are still confidential:
    - If the social media account messages belong to the right owner.
- Current studies focus on semantical analysis or stylometric analysis
- Our current focus is on the author identification through character analysis
An IRC client is required to connect to IRC servers, where numerous chat channels are open for users to join.

Any chat sent by users are immediately broadcast to all in the same channel.

A protocol that facilitates real-time text communications.

IRC has been extensively used by hacker over years.
Architecture

IRC Channel → Autonomic Monitoring Bot → IRC Message Storage → Personality Feature Extraction → Personality Feature Storage

New IRC Messages → Personality Feature Extraction → Author Identification Model → Result of Author Identification on the New Messages

Learning Unit → Personality Training Samples
Author Characterization

```
{
  "processed_language": "en",
  "word_count": 19469,
  "personality": [
    {
      "trait_id": "big5_openness",
      "name": "Openness",
      "category": "personality",
      "percentile": 0.820061335253919,
      "children": [
        {
          "trait_id": "facet_adventurousness",
          "name": "Adventurousness",
          "category": "personality",
          "percentile": 0.7201522774034973
        },
        {
          "trait_id": "facet_artistic_interests",
          "name": "Artistic interests",
          "category": "personality",
          "percentile": 0.5329473652969766
        }
      ]
    }
  ]
}
```
Experimental Results

Author Identification Accuracy in #anonops

Accuracy

100.00%
95.00%
90.00%
85.00%
80.00%
75.00%

Number of authors

5
10
20
30

1-NN
3-NN
5-NN
SVM
Experimental Results

Author Identification Accuracy in #politics

Accuracy

Number of authors

1-NN  3-NN  5-NN  SVM
Experimental Results

Author Identification Accuracy in #2600

Accuracy vs. Number of authors:
- 1-NN
- 3-NN
- 5-NN
- SVM
Experimental Results

Author Identification Accuracy in computer

Accuracy

Number of authors

5 10 20 30

1-NN 3-NN 5-NN SVM
Milestones and Deliverables for 2018

Milestones for the current proposed year:

- Q1: Building the framework, including registration & licensing system together with footprint
- Q2-Q3: Creating speech recognition & voice based service oriented commands and creating thin-pads to be used for the personal assistant
- Q4: Building proof-of-concept environment

Deliverables for the current proposed year:

1. Expanding framework;
2. Improving the personal assistant capabilities and connecting to the services;
3. Various services;
4. Public reports or publications.
Potential Member Company Benefits

Expected benefits to members and other CAC projects include:

• Applying personal analysis for better understanding individuals
  • This can help to a wide range of members

• Help community through services such as
  • Providing security in validation of the user
  • Understanding the malicious users/groups
Please take a moment to fill out your L.I.F.E. forms.

http://www.iucrc.com
Select “Cloud and Autonomic Computing Center” then select “IAB” role.

What do you like about this project?
What would you change?
(Please include all relevant feedback.)