Exploring Dark Networks: From the Surface Web to the Dark Web

Hsinchun Chen, Ph.D., University of Arizona
Panel 3: Multi-Level, High-Dimensional Evolving and Emerging Networks (ML-HD-EEN)

October 12, 2017
“Leveraging Advances in Social Network Thinking for National Security: A Workshop”

Acknowledgements: NSF, DHS, DOJ, DOD
My Background

• Academic: Data/computational scientist; data/text/web mining, visualization; applied AI for security and health analytics

• Applications: Security analytics; dark networks

• Projects: COPLINK (1997-2009, gang/narcotic networks); Dark Web (2001-present, extremist/terrorist networks); Hacker Web (2009-present)

• SBE collaborators: M. Sageman, R. Breiger, T. Holt

• Agency collaborators: TPD, PPD, FBI, CIA, NSA, DHS (NSF, DOD)
Springer, 2006

Intelligence and Security Informatics for International Security
Information Sharing and Data Mining
Hsinchun Chen

Springer, 2012

Dark Web
Exploring and Data Mining the Dark Side of the Web
Hsinchun Chen
A Vehicle to Watch via its Networks?

Shape Indicates Object Type
- circles are people
- rectangles are vehicles

Color Denotes Activity History
- Red: Gang related
- Yellow: Violent crimes
- Blue: Narcotics crimes
- Green: Violent & Narcotics

Larger Size Indicates higher levels of activity

Border Crossing Plates are outlined in Red
COPLINK Identity Resolution and Criminal Network Analysis (DHS)

**Cross-jurisdictional Information Sharing/Collaboration**

- Arizona IDMatcher
- Law-enforcement Data: Arizona (AZ), California (CA), Texas (TX)
- Border Crossing Data: Arizona (AZ), California (CA), Texas (TX)
- CAN Visualizer

**Identity Resolution**
- Detect false and deceptive identities across jurisdictions using a probabilistic naive Bayes based resolution system.

**Criminal Network Analysis**
- **High-risk Vehicle Identification**
  - Identify high-risk vehicles using association techniques like mutual information using border crossing and law enforcement data.
- **Criminal Link Prediction**
  - Predict interaction between individuals and vehicles using link prediction techniques to identify high-risk border crossers.
- **Suspect Traffic Burst Detection**
  - Detect real-time anomalies and threats in border traffic using Markov switching and other models.

*Only the grayed datasets are available to the AI Lab*
Dark Web Overview

- Dark Web: Terrorists’ and cyber criminals’ use of the Internet
- Collection: Web sites, forums, blogs, YouTube, etc.
- 20 TBs in size, with close to 10B pages/files/messages (the entire LOC collection: 15 TBs)
Dark Web Forum Crawler System: Probing the Hidden Web (Proxy, TOR)
CyberGate for Social Media Analytics: Ideational, Textual and Interpersonal Information
### Arabic Writeprint Feature Set: Online Authorship Analysis

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Lexical</th>
<th>Syntactic</th>
<th>Structural</th>
<th>Content Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(79)</td>
<td>(262)</td>
<td>(62)</td>
<td>(15)</td>
</tr>
<tr>
<td>Lexical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(48)</td>
<td>(12)</td>
<td>(14)</td>
<td></td>
</tr>
<tr>
<td>Char-Based</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(48)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word-Based</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(31)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Punctuation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(200)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhetorical Words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(50)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word Structure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical Structure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(11)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race/Nationally</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Violence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Feature Details

- **Lexical**

- **Syntactic**

- **Structural**

- **Content Specific**
  - (15) Race/Nationally, (4) Violence.
Arabic Feature Extraction Component

1. Incoming Message
   - منتدى الأنصار الإسلامي

2. Elongation Filter
   - Filtered Message
   - منتدى الأنصار الإسلامي
   - Count +1
   - Degree + 5

3. Root Clustering Algorithm
   - Root Dictionary
   - Generic Feature Extractor
   - Similarity Scores (SC)
     - انقلز 0.54
     - منقلز 0.21
     - انقلر 0.31
   - max(SC)+1
   - All Remaining Features Values

4. Feature Set
CyberGate System Design: Writeprints
Author Writeprints

Anonymous Messages

10 messages

10 messages
AZ Forum Portal

• 13M messages (340K members) across 29 major Jihadi forums in English, Arabic, French, German and Russian (VBulletin)

• Linking members over time
Moving Toward Black Hat Research in Information Systems Security: An Editorial Introduction to the Special Issue

By M. Adam Mahmoud
University of Texas at El Paso
mamhmod@utep.edu

Mikko Siponen
University of Oulu, Finland
mikko.siponen@oulu.fi

Detmar Straub
Georgia State University
dstraub@gasu.edu

H. Raghav Rao

Black Hats Versus White Hats
Versus Grey Hats

What exactly is this white hat versus the black hat dichotomy? When making movies about the Old American West, filmmakers made a symbolic distinction at times between the good guys, wearing white hats, and the bad guys, wearing black hats. If, for the sake of our basic theme, we can adopt this distinction momentarily, we would like to go on to assert that the information systems field is heavily over-emphasizing research on white hats to the detriment of studies on black hats. It is easy to see how this would, quite naturally, occur. Scholars have better access to white hats,

When hackers talk, this research team listens

Discovery

Online conversations help fill critical gap in cybersecurity knowledge about attackers' motivations, possible targets

October 8, 2015

Hans-Hyung Chun leads a research project that explores the motivations of cyberthieves.

Credit and Larger Version

National Science Foundation
WHERE DISCOVERIES BEGIN

FUNDING AWARDS DISCOVERIES NEWS PUBLICATIONS STATISTICS ABOUT NSF FASTLANE

Funding

1. DIRECTorate for Computer & Information Science & Engineering

Secure and Trustworthy Cyberspace (SaTC)

Division of Graduate Education

CyberCorps(R) Scholarship for Service (SFS)

Email Print Share

Email Print Share

Email Print Share

Email Print Share
Hacker Assets Portal V2.0 – Overview

(a) Home page, linking to (b & c) Assets, (d) Dashboard, and (e) Malware Families:

(b) Assets page, linking to Source Code and Attachments

(c) Source Code page; sortable by asset name, exploit type, date, etc.

(d) Dashboard for drill-down analysis of hackers & assets over time

(e) Malware Families, for depicting relationships among assets over time (Crypter Family shown)
1. Filtering on 2014, when BlackPOS was posted, shows assets and threat actors at that time.

2. Filtering the actor who posted BlackPOS reveals that he posts other bank exploits (e.g., Zeus).
   • Provides intelligence on which hacker to monitor.
1. Filtering on a specific time point (highest peak):
2. Filtering on a specific asset (crypters, a key technology for Ransomware)
3. Filtering a specific crypter author (Cracksman) shows the trends and types of assets he posted.
Selected Challenges for ML-HD-EEN in Dark Networks

- Identifying data Sources: availability (data stovepipes, data integration; RMS, RDBMS); web OSINT (surface web, deep web, dark web; TOR, ICT); data types (structured vs. unstructured; multi-lingual, multi-media; source code, attachment, tutorial), data biases (noise, deception, adversarial; vigilante, honeypot, APTs)

- Recognizing nodes: levels/dimensions (who/what/where/when/why/how); entity extraction and recognition (identity resolution, web authorship analysis, writeprint)

- Establishing links: linked by associations (labeled links, probabilistic links); linked by time/space (same-time-same-place; border crossing, hotspot); linked by conversations (linguistic cues and styles; ICT, forums)

- Analyzing network patterns: (many SNA techniques)

- Tracking changes over time: stream data collection & mining (update & alert; anomaly detection, concept drift; emerging cyber threats and DarkNet Markets)
Selected Solutions & Directions for ML-HD EEN

• Comprehensive & timely OSINT data collection: from the surface web to the dark web; across level/dimension, over time

• Data integration and SNA extraction: AI assisted entity/relationship recognition/integration; across level/dimension, over time

• Methodological foundations: dark networks, hidden networks; noise, deception, adversarial intent

• Data analytics: advanced social media analytics, stream data mining, adversarial machine learning, BIG DATA analytics; across level/dimension, over time
For questions and comments

hchen@eller.Arizona.edu
http://ai.Arizona.edu