“Cybersecurity Big Data and Analytics Sharing”

• Big Data ➔ Big Data Breach in Cybersecurity!

• Hsinchun Chen, UA, Conference/Workshop Chair, time keeper
• 2 sessions, 10-12 mins for each speaker; Q/A/contribution from audience after session (part of an NSF report)
• Session I: Bhavani Thuraisingham, UT Dallas, malware analysis; Latifur Khan, UTD, stream data analytics; Victor Benjamin, Arizona State U, blockchains for cybersecurity
• Session II: Resha Shenandoah, UA, Data Infrastructure Building Block (DIBBs) for security data; Sagar Samtani, UA, DIBBs tools, Hacker Assets Portal; Weifeng Li, UA, hacker underground economy (UA/Eller/MIS AI Lab)

Acknowledgement: National Science Foundation under Grant Number ACI-1443019 (DIBBs) & DGS-1719477 (SFS/SaTc)
Session Break Questions to Consider

• Questions and comments relating to session talks: What data or tools do you consider to be most useful for you and why? Other comments?

• Questions and comments relating to workshop in general: What additional data or tools do you wish to have and why? Other comments?

• Speaker slides/content and audience responses will be summarized in an NSF Workshop Report for distribution. (Please contact rshenandoah@email.Arizona.edu.)

Acknowledgement: National Science Foundation under Grant Number ACI-1443019 (DIBBs) & DGS-1719477 (SFS/SaTc)
Malware Data Collection & Analysis Using Big Data Tools

Cyber Security Research & Education Institute
The University of Texas at Dallas
Ramkumar Paranthaman
Dr. Bhavani Thuraisingham
Agenda

- Introduction
- Malware Data Collection
  - Malware Data Types
  - Malware Dataset Classification
  - Malware Collection Statistics
- Malware Analysis
  - Feature Extraction
  - Feature Selection
  - Train ML models
  - Results
Introduction

- This NSF-funded Data Infrastructure Building Blocks (DIBBs) project is intended to address a large gap in the availability of open source research data for researchers in ISI.
- The University of Arizona Artificial Intelligence Lab and its partners, the University of Virginia, The University of Texas at Dallas, Drexel University, and the University of Utah were to collect a significant archive of data and analysis tools to serve the ISI community.
- http://www.azsecure-data.org/
Data Collection - Repositories

- **Classified datasets**
  - Academic research projects
  - Security research corporations
- **Unclassified datasets**
  - Public malware datasets
  - Non-corporate research group malware datasets
- **Malware collections**
  - Independent collections of malware data
  - Malware sharing through forums
Malware Collection - Statistics

- Number of Classified Datasets: 25 (circa 230 GB)
- Number of Unclassified Datasets: 16 (circa 26 GB)
- Independent Datasets Gathered: 3 (circa 2 GB)
- Total Size of Malware Datasets: circa 250 GB
Malware Detection Framework

- **Objective**
  - Develop a malware detection framework using static analysis approach by employing Big Data tools and machine learning techniques

- **Implementation Steps**
  - Feature Extraction
  - Feature Selection
  - Training
  - Classification (detection)
System Workflow

Training data → Feature Extraction → Feature Selection → Pruned Training Data

Test Data → Feature Extraction → Pruned Test Data

Pruned Training Data → Training (SVM, RF) → Model

Predict → Predicted Output

RF  Random Forest
SVM  Support Vector Machine
Feature Extraction - Map-Reduce Workflow
Feature Extraction - Map-Reduce Workflow
Feature Selection & Training

- **SELECTION**
  - Compute information gain for each feature
  - Select features whose information gain is above a threshold value

- **TRAINING - LEARNER MODELS**
  - Support Vector Machine (radial basis kernel)
  - Random Forest (J48 tree)
Inference

- **Input Dataset**
  - Size - 200 GB
  - Type - Executables, DLLs, Hexcode dumps

- **Observation**
  - Process time - 29.37 mins
  - Hadoop - though highly scalable, lacks performance due to high I/O usage (especially for large volume datasets)

- **Mitigation**
  - Use Apache Spark, distributed in-memory computing framework to improve performance
RESULTS

- Feature Extraction

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Byte 4-grams</td>
<td>95, 608, 217</td>
</tr>
<tr>
<td>Assembly 4-grams</td>
<td>419,888</td>
</tr>
<tr>
<td>DLL imports</td>
<td>26, 785</td>
</tr>
<tr>
<td>Opcode frequencies</td>
<td>82</td>
</tr>
</tbody>
</table>

- Feature Selection

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Byte 4-grams</td>
<td>46, 317</td>
</tr>
<tr>
<td>Assembly 4-grams</td>
<td>4, 309</td>
</tr>
<tr>
<td>DLL imports</td>
<td>65</td>
</tr>
<tr>
<td>Opcode frequencies</td>
<td>Not pruned</td>
</tr>
</tbody>
</table>

- Classification

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>96.31%</td>
</tr>
<tr>
<td>Support Vector machine</td>
<td>95.05%</td>
</tr>
</tbody>
</table>

- Processing time - 16.42 minutes

- Source code
  [https://github.com/helloram52/detectmalware](https://github.com/helloram52/detectmalware)
Trends and Perspectives in Big Data Research and Application

Latifur Khan, PhD
Professor
Department of Computer Science
University of Texas at Dallas, lkhan@utdallas.edu
Big Data: Issues

• Real Time
  • Data Processing Overhead needs to be minimized
    • Large Volume of Data needs to be consumed
  • Analytics
    • Response needs to be in real time.
    • Example: Real Time Anomaly Detection*
      • False Alarm may increase

• Scalable Analytics
  • Many Typical Algorithms Suitable for In-Memory processing
  • Demands Distributed Processing

Big Data: Solution

• **Real Time Processing**
  • Tool: Apache Spark, Storm, S4, Flink

• **Real Time Analytics**
  • SAMOA

• **Scalable Analytics**
  • Tool: Spark’s Machine Learning Library (MLlib), Mahout etc.

• **Covers Basic Analytics Algorithms**

• **Advanced Algorithms (Relational Learning) are missing**


  • *Ahsanul Haque, Zhuoyi Wang, Swarup Chandra, Yupeng Gao, Latifur Khan, Charu Aggarwal, Sampling-based distributed Kernel mean matching using spark. BigData 2016: 462-471*
Big Data: Current & Future

- Stream Mining*
  - Update Learner Continuously
- Analytics
  - Supervised Learning (Ground Truth is required)
  - Labeling of Data is Problematic
  - Active Learning+

+Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani M. Thuraisingham: A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data. ICDM 2008: 929-934
Challenges: Fixed Chunk Size

Concept Drifts

Chunk size too large – Delayed reaction

Chunk size too small – Performance issue

Correct

Wrong
Solution: Adaptive Chunk Size

Concept Drifts

- Correct
- Wrong

Adaptive Chunk Size
Existing dynamic sliding window techniques

monitor error rate of the classifier.

Update classifier if starts to show bad performance.

fully supervised, which is not feasible in case of real-world data streams.

Adaptive Chunk - Unsupervised

Haque et al. [1][2]

Input → Prediction using Ensemble → Predicted Class

Classifier Confidence

Distribution Before → Update Classifier & Shrink Window
Distribution After

Yes → Chang e
No → Grow Window

Big Data: Current & Future

- Stream Mining*
  - IOT Big Stream Mining
  - Security:
    - Encrypted Stream Traffic Analysis
      - Website Fingerprinting

Application: Encrypted Traffic Fingerprinting

Al-Naami et al. [1][2]

• Traffic Fingerprinting (TFP) is a Traffic Analysis (TA) attack that threatens web/app navigation privacy.
• TFP allows attackers to learn information about a website/app accessed by the user, by recognizing patterns in traffic.
• Examples: Website Fingerprinting

Application: Real-time Political Actor Detection Over Textual Political Stream

Challenges

- Same actor with multiple alias names
- Identify novel actor along with roles
- Existing political actor’s role changes over time
- Processing high volume of news articles across the world

Blockchains for Cybersecurity Research

Victor Benjamin, Ph.D.
Assistant Professor, Department of Information Systems
Co-Director, Actionable Analytics Lab
Introduction - Problem Context

• Industry thinks cybersecurity data sharing is good
  – Business-to-Business sharing (e.g., supply chains)
  – Business-to-Government (e.g., incident sharing)

• In reality, common reluctance to share data
  – Liability
  – Accessibility, transparency, and data ownership
  – Sharing platform focus, usefulness, and usability
What can fix this?

• Need for platform that supports consortiums
  – Encourages community building
  – Can cater to special interest groups
    • E.g., Maritime Information Sharing and Analysis Center
A Path Forward

• Blockchains, the technology behind Bitcoin
  – First work on crypto-secured chain of blocks in 1991
  – First “modern” conceptualization of Blockchain in 2008

• Peer-to-peer networks
  – Managed autonomously
  – Highly configurable
Bitcoins – A Quick Primer

1. A wants to send money to B
2. The transaction is represented online as a 'block'
3. The block is broadcast to every party in the network

Those in the network approve the transaction is valid
The block then can be added to the chain, which provides an indelible and transparent record of transactions
The money moves from A to B
Blockchain Characteristics

- A distributed computing infrastructure offering:
  - Decentralized
  - Resiliency
  - Immutability
  - Security
  - Privacy

- Qualities for a cybersecurity data sharing platform
Blockchain for Cybersecurity Data

1. A wants to send money to B
2. The transaction is represented online as a ‘block’
3. The block is broadcast to every party in the network
4. Those in the network approve the transaction is valid
5. The block then can be added to the chain, which provides an indelible and transparent record of transactions
6. The money moves from A to B
“All people everywhere should have free energy sources.”

[...] “Electric Power is everywhere present in unlimited quantities and can drive the world’s machinery without the need for coal, oil or gas.”

~ Nikola Tesla (1856-1943)

EU Sustainable Energy Week HACKED!!!
#EUSEW2014

more than 10,000 accounts from companies and governments
function, email, telephone, password

Worldbank, Bayer, ExxonMobil, Enel, Edf, GE, Shell, BP, Eni, Nokia, Intel... and many others.

operation initiative to encourage real #sustainable #energy

Hacked List:
https://docs.zoho.com/file/egrja2d5495484a724a05a2495e9e73f81dcc (filetype .csv)
https://docs.zoho.com/file/egrja900ddc3ea08942d992eee71fc6b9f024 (filetype .ods)

DOWNLOAD:
https://cdn.anonfiles.com/1403546269644.txt

“Operation Green Rights”
Use Case: Threat Analytics

Cyber-Threat Data Collection
- Collection Manager
- Transformed into ‘Blocks’ for Storage
- Forums
- Markets
- Partner Data

Blockchain Infrastructure
- User Interface
- Application Layer
- Blockchain Back-end
- Blockchain Data Access API

Threat Analytics
- Analytics Center
- Language Modeling
- Market Analysis
- Threat & Target Analysis
Use Case: Cyber-physical Security

Modeling and Analysis of Attacker-Defender Strategies
- Game Theoretic Modeling
- Emerging Threat Intelligence
- Cyber-attack Attribution

Securing Cyber-physical Systems Infrastructure
- Commercial Off-the-Shelf IoT Security Analysis
- Attributed-based Moving Target Defense
- Blockchain-enabled Distributed Monitoring

Protection of Cyber-physical Devices
- Monitoring Devices and Interactions
- Device Access Control
- Physical Interface Security

Application Areas
- External Interfaces
- Insider Threats
- Cross-organizational

Networks
- SCADA Systems
- Smart Infrastructure

IoT Devices
- Embedded Systems
- Sensors

W.P. Carey School of Business
Arizona State University
Conclusion

• Platforms must be built with stakeholders in mind
• Blockchains offer a unique opportunity
• General take-away: think outside the box
  – Hackers do it
  – Security researchers and practitioners need to as well
Thanks!

• Ongoing exploration and proto-typing

• Interested? Contact me Victor.Benjamin@asu.edu
Resha Shenandoah

Digital Archivist

Project Manager, Data Infrastructure Building Blocks for Intelligence and Security Informatics (DIBBs-ISI)
University of Arizona Artificial Intelligence Lab

Pl: Dr. Hsinchun Chen, University of Arizona. Co-PIs: Dr. Mark Patton and Cathy Larson, University of Arizona. Project Partners: Dr. Ahmed Abbasi, University of Virginia; Dr. Paul Hu, University of Utah; Dr. Bhavani Thurasingham, University of Texas at Dallas; Dr. Chris Yang, Drexel University.

This material is based in part upon work supported by the National Science Foundation under Grant Number ACI-1443019.
DIBBs-ISI: azsecure-data.org

14 Collections, 200+ GB total

<table>
<thead>
<tr>
<th>Websites:</th>
<th>Phishing</th>
<th>171,360</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>US Patriot, Hate, Militia 2009</td>
<td>74 identified by SPLC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>133 linked</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Forums:</th>
<th>Geo Web</th>
<th>65</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dark Web</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Hacker</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Chinese underground economy</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network Traffic:</th>
<th>4 collections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malware Instances:</td>
<td>25,118 unique instances from 1 collection</td>
</tr>
</tbody>
</table>

Also collections containing chat logs and international news.

**Languages:**
- Arabic, Chinese, English, French, German, Indonesian, Pashto, Russian, Urdu

**File types:**
- arff, asp, binetflow, cfm, class, css, csv, exe, ghc, html, java, mpg, pcap, pdf, php, rar, sql, swf, txt, wd3, webarchive, wma, wmv, xlsx
Between August 2016 and March 2017:

- 1,404 GB of data downloaded
- 17,190 file requests
- 51 distinct countries/regions originating requests

Most requested collection: PhishMonger

- 14,551 file requests
azsecure-data.org: PhishMonger

- Invokes the PhishTank API hourly
  - Indexes online, valid phishing sites
  - Typically 25,000 to 50,000 sites per request
  - Updated hourly
- Identifies newly added phish URLs
- Fetches new phishing websites
- Saves data

Targeted Brands with 200+ Sites

IEEE-ISI, September 2016

Targeted Brand

Frequency
Leverages exclusively open source software:
- Ubuntu Linux, GNU Wget, Filezilla Server FTP

Coded in Python 3.5
- Harnesses the Twisted library for time based scheduling

Runs on Amazon Web Services (AWS) Elastic Compute Cloud (EC2)

Additional statistical scripts written in R

Most common file types include: png, html, jpg gif, js, css, ttf, svg, ico, woff

Contact:
- Ahmed Abbasi, abbasi@comm.virginia.edu
- David G. Dobolyi, dd2es@comm.virginia.edu
DSpace
Metadata
Search or Browse
OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting
Brings data out of silos
Persistent identifiers – DOI, Orchid
Built-in analytics
Data Preservation

Alice Through the Looking Glass, ca 1871

HRC Clay Tablet. Sumerian. Ca 2400 BCE
http://www.hrc.utexas.edu/educator/modules/gutenberg/books/early/
One-on-one outreach and training is highly effective, but not efficient or scalable.
Platform for Cybersecurity Big Data

• Resource for research data management
• Automate metadata creation
• Pipeline
  • Data management for research
  • Data sharing

• Continued user interest improves chances of sustaining cyberinfrastructure

• Must meet user needs

Resha Shenandoah – rshenandoah@email.arizona.edu
azsecure-data.org
DIBBs Tool Inventory for ISI Research

Sagar Samtani, Shuo Yu, Weifeng Li, Hongyi Zhu, Resha Shenandoah, Hsinchun Chen
Artificial Intelligence Lab, The University of Arizona
March 31, 2017

*This material is based upon work supported by the National Science Foundation under Grant No. NSF ACI-1443019*
Introduction

• Security researchers may face steep learning curves when attempting to identify tools that can aid them in developing valuable security insights from data sets.

• These slides aim to reflect some tools in the data analytics landscape that have been used in the AI Lab’s past security informatics research.

• We present an inventory of tools into three major sections based on a traditional data analytics pipeline:
  • Collection and storage tools
  • Pre-processing and analytics tools
  • Visualization tools

• We also select a set of ISI papers to show how the tools can be used together to facilitate research.
Collection and Storage Tools

• The collection and storage component of relevant data is the first stage in typical data analytics exercises.

• Data collection aims to:
  • Identify and capture relevant fields of data from a specific source (e.g., web forums, Twitter, etc.)
  • Index and store it in a database or some other format which can be retrieved and used for pre-processing and further analytics.

• The collection process comprises three steps to pull from the online sources into the database: extract, transform, and load (ETL).
  • Table 1 summarizes tools to perform such tasks.
# Collection Process: ETL

<table>
<thead>
<tr>
<th>Collection Stage</th>
<th>Description</th>
<th>Category</th>
<th>Tool Name</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Extract</strong></td>
<td>Extracting data from their sources (e.g., websites, API’s)</td>
<td>Spiders</td>
<td>Offline Explorer</td>
<td>GUI for scheduling various crawling projects</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cURL</td>
<td>Offers proxy support, user authentication, etc.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Wget</td>
<td>Recursive download, conversion of links</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Packages for Customized Spiders</td>
<td>HtmlUnit</td>
<td>A headless web browser written in Java</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Serenium</td>
<td>A browser automation library in Python</td>
</tr>
<tr>
<td><strong>Transform</strong></td>
<td>Transforming raw data into target data elements</td>
<td>Transformation</td>
<td>Regex</td>
<td>General string pattern matching</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>JSoup</td>
<td>Java library for parsing HTML</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BeautifulSoup</td>
<td>Python package for parsing HTML and XML</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>urllib</td>
<td>High-level interface for fetching data across the Web</td>
</tr>
<tr>
<td><strong>Load</strong></td>
<td>Loading data into data warehouse</td>
<td>Databases</td>
<td>MySQL</td>
<td>Widely used open-source RDBMS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MS SQL Server</td>
<td>Commercial RDBMS by Microsoft</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Oracle Database</td>
<td>Commercial RDBMS by Oracle</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Apache HBase</td>
<td>Open-source, distributed, NoSQL DBMS on top of Hadoop</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Apache Hive</td>
<td>Open-source data warehouse infrastructure on top of Hadoop</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MongoDB</td>
<td>Open-source NoSQL DBMS. Uses JSON-like documents with schemas</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Apache Lucene</td>
<td>High-performance, full-featured text search engine library in Java</td>
</tr>
</tbody>
</table>

Table 1. Extraction, Transformation and Loading Tools
Pre-Processing and Analytics Tools

- Collected data needs to be pre-processed and transformed (cleaning, normalizing, transforming, tokenizing, etc.) prior to analysis.
  - Often consumes the majority (70-75%) of the time in data analytic projects.

- Past security analytics have used dozens of techniques after pre-processing, ranging from summary statistics to complex algorithms (e.g., deep learning).

- Many common data and text mining algorithms/applications are bundled into single packages (e.g., WEKA, Natural Language Toolkit (NLTK)).
  - Other analytics offered in specialized packages (e.g., hidden Markov models (HMM))

- Tables 2 and 3 summarize various pre-processing and analytical tools.
# Pre-Processing and Analytics Tools

<table>
<thead>
<tr>
<th>Category</th>
<th>Tool Name</th>
<th>Programming Language</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Data Mining</td>
<td>WEKA</td>
<td>Java, GUI</td>
<td>One-stop tools that cover common pre-processing, classification, and clustering algorithms. RapidMiner and WEKA can be used independently without a specific programming language.</td>
</tr>
<tr>
<td></td>
<td>Scikit-learn</td>
<td>Python</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RapidMiner</td>
<td>GUI</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>R</td>
<td>A widely used programming language and software environment for statistical computing and graphics.</td>
</tr>
<tr>
<td>General Text Mining</td>
<td>Natural Language Toolkit (NLTK)</td>
<td>Python</td>
<td>One-stop tools that cover word/sentence tokenization, POS tagging, parsing, chunking, named entity recognition, etc. NLTK has interfaces to call Stanford NLP tools.</td>
</tr>
<tr>
<td></td>
<td>Stanford CoreNLP</td>
<td>Java</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Apache OpenNLP</td>
<td>Java</td>
<td></td>
</tr>
<tr>
<td>Hidden Markov Models (HMM)</td>
<td>hmmlearn</td>
<td>Python</td>
<td>General HMM package</td>
</tr>
<tr>
<td></td>
<td>NLTK</td>
<td>Python</td>
<td>Specialized in POS tagging</td>
</tr>
<tr>
<td>Conditional Random Fields (CRF)</td>
<td>Stanford NER CRF</td>
<td>Java</td>
<td>CRF implementation for named entity recognition (NER)</td>
</tr>
<tr>
<td></td>
<td>CRF++</td>
<td>C++</td>
<td>General CRF package</td>
</tr>
<tr>
<td></td>
<td>NLTK</td>
<td>Python</td>
<td>Specialized in POS tagging</td>
</tr>
<tr>
<td>Latent Dirichlet Allocation (LDA)</td>
<td>Mallet</td>
<td>Java</td>
<td>Command line based tool for standard LDA</td>
</tr>
<tr>
<td></td>
<td>Stanford Topic Modelling Toolbox</td>
<td>GUI</td>
<td>GUI based tool that supports LDA, labelled LDA, partially labelled LDA, and calculating perplexity. Can also perform temporal LDA</td>
</tr>
<tr>
<td></td>
<td>Gensim</td>
<td>Python</td>
<td>Perform latent semantic analysis (LSA) and LDA in Python</td>
</tr>
</tbody>
</table>

Table 2. General and Specialized Data and Text Mining Tools
## Pre-Processing and Analytics Tools

### Table 3. General and Specialized Data and Text Mining Tools (cont’d)

<table>
<thead>
<tr>
<th>Category</th>
<th>Tool Name</th>
<th>Programming Language</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Network Analysis (SNA)</td>
<td>UCINET</td>
<td>GUI</td>
<td>Licensed software (minimum $40) that can handle medium sized networks (2 millions nodes max)</td>
</tr>
<tr>
<td></td>
<td>Gephi</td>
<td>GUI</td>
<td>Open source GUI based software that can handle larger data sizes than UCINET. Can read directly from databases</td>
</tr>
<tr>
<td></td>
<td>NetworkX</td>
<td>Python</td>
<td>Python based network analysis tools. Can read from a variety of data sources. Allows for significant customization compared to other tools</td>
</tr>
<tr>
<td>Ontologies</td>
<td>WordNet</td>
<td>-</td>
<td>English lexical database grouped into synonyms</td>
</tr>
<tr>
<td></td>
<td>SentiWordNet</td>
<td>-</td>
<td>Tagged WordNet with positivity, negativity, and neutrality for opinion mining</td>
</tr>
<tr>
<td>Word2vec</td>
<td>Gensim</td>
<td>Python, C</td>
<td>A two-layer neural net that processes text. Outputs a set of vectors: feature vectors for words in that corpus. Turns text into a numerical form for deep nets.</td>
</tr>
<tr>
<td></td>
<td>DL4J</td>
<td>Java, Scala</td>
<td></td>
</tr>
<tr>
<td>Deep Learning</td>
<td>Keras</td>
<td>Python</td>
<td>High-level neural networks library running on top of either TensorFlow or Theano. Recommended for fast experimentation.</td>
</tr>
<tr>
<td></td>
<td>TensorFlow</td>
<td>Python, C++</td>
<td>Two different low-level implementations for deep learning models</td>
</tr>
<tr>
<td></td>
<td>Theano</td>
<td>Python</td>
<td></td>
</tr>
</tbody>
</table>
Visualization Tools

• The final stage in the data often incorporates a visualization component.

• Desktop software (table 4) provide turnkey solutions to manage, connect, pivot data and render predefined types of visualizations in a GUI.

• For better customizability, lightweight toolkits, packages, and online services can be implemented along with analytical scripts (table 5).

<table>
<thead>
<tr>
<th>Tool Name</th>
<th>Cost</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Excel</td>
<td>License required</td>
<td>Excel supports charts, graphs, generated from specified groups of cells. Excel 2010 and later support Pivot Table.</td>
</tr>
<tr>
<td>Tableau</td>
<td>Free education license</td>
<td>Generates graph types that can be combined into dashboards and shared over the internet.</td>
</tr>
<tr>
<td>ParaView</td>
<td>Free, open-source</td>
<td>Developed to analyze extremely large datasets using distributed memory computing resources.</td>
</tr>
</tbody>
</table>

Table 4. Desktop Visualization Software
# Visualization Tools

<table>
<thead>
<tr>
<th>Category</th>
<th>Tool Name</th>
<th>Programming Language</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Data Visualization Toolkits</td>
<td>Visualization Toolkit (VTK)</td>
<td>C++, Python, Java</td>
<td>General tools enabling users to customize their visualization components (e.g., point, line, axes, legends, layout, color coding) programmatically.</td>
</tr>
<tr>
<td></td>
<td>OpenFrameworks (OF)</td>
<td>C++</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matplotlib</td>
<td>Python</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ggplot2</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Processing</td>
<td>Java, Python, JS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Seaborn</td>
<td>Python</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pandas</td>
<td>Python</td>
<td></td>
</tr>
<tr>
<td>Word Cloud</td>
<td>Wordle</td>
<td>Online, JS</td>
<td>Word cloud is a graphical representation of word frequencies. It can be used to visualize most frequently used keywords in the corpus.</td>
</tr>
<tr>
<td>Geo-map Tools</td>
<td>Mapbox</td>
<td>Online, JS</td>
<td>When location data (e.g. state, zipcode, latitude and longitude) is available, these geo-map tools can help you layout the data onto a map and generate visualizations such as color map, flow maps, etc.</td>
</tr>
<tr>
<td></td>
<td>geoplotlib</td>
<td>Python</td>
<td></td>
</tr>
<tr>
<td></td>
<td>choroplethr</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>Network Visualization Tools</td>
<td>Gephi</td>
<td>GUI, Java</td>
<td>Network visualization tools can visualize the relationship between data attributes or different data sources. The built in layout algorithms automatically generate visually pleasing graphs.</td>
</tr>
<tr>
<td></td>
<td>networkx</td>
<td>Python</td>
<td></td>
</tr>
<tr>
<td></td>
<td>graph-tool</td>
<td>Python</td>
<td></td>
</tr>
<tr>
<td></td>
<td>igraph</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>Color Selection (Aesthetic)</td>
<td>Color Brewer 2</td>
<td>Online</td>
<td>These color selection tools helps to improve the aesthetic of the visualization. They also provide safe color selections for web presenting, printing, color-blind cases.</td>
</tr>
<tr>
<td></td>
<td>Palettable</td>
<td>Python</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RColorBrewer</td>
<td>R</td>
<td></td>
</tr>
</tbody>
</table>

**Table 5. Lightweight Toolkits, Packages, and Online Services**
Example ISI Papers

- To show the research context of applying the listed tools, we reviewed over 100 research papers from past ISI conferences and workshops.
  - 56 papers from IEEE ISI 2016
  - 47 from IEEE ISI 2015
  - 8 from FOSINT-SI 2016
  - 10 from ISI-ICDM 2015
- We selected representative papers to show how those tools can be used together to support and facilitate research.
  - References are attached at the end.
## Example ISI Papers

<table>
<thead>
<tr>
<th>Paper</th>
<th>Collection and Storage</th>
<th>Pre-Processing and Analytics</th>
<th>Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samtani et al. (2016)</td>
<td>Offline Explorer, MySQL, Regex</td>
<td>RapidMiner, Stanford Topic Modelling Toolbox</td>
<td>Tableau, D3.js</td>
</tr>
<tr>
<td>Grisham et al. (2016)</td>
<td>Selenium, MySQL</td>
<td>Stanford Topic Modelling Toolbox</td>
<td>-</td>
</tr>
<tr>
<td>Benjamin &amp; Chen (2016)</td>
<td>Offline Explorer, MySQL, Regex</td>
<td>Word2vec</td>
<td>-</td>
</tr>
<tr>
<td>Benjamin &amp; Chen (2014)</td>
<td>IRC Bots</td>
<td>WEKA</td>
<td>-</td>
</tr>
<tr>
<td>Samtani &amp; Chen (2016)</td>
<td>Offline Explorer, MySQL, Regex</td>
<td>Gephi</td>
<td>Gephi</td>
</tr>
<tr>
<td>Solaimani et al. (2016)</td>
<td>MongoDB</td>
<td>CoreNLP, WordNet</td>
<td>-</td>
</tr>
<tr>
<td>Dobolyi &amp; Abbasi (2016)</td>
<td>PhishTank API, Wget</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>Park et al. (2016)</td>
<td>SQLite</td>
<td>Apache OpenNLP</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 6. Example ISI Papers
References


• Solaimani, M., Salam, S., Mustafa, A. M., Khan, L., Brandt, P. T., & Thuraisingh, B. (2016, September). Near real-time atrocity event coding. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 139-144). IEEE.
AZSecure Hacker Assets Portal: Enhancing Cybersecurity Education

Sagar Samtani, Kory Chinn, Cathy Larson, Hsinchun Chen
Artificial Intelligence Lab, The University of Arizona
March 31, 2017

*This material is based upon work supported by the National Science Foundation under Grant No. NSF DUE-1303362 (SFS) and NSF SES-1314631 (SaTC).*
AZSecure Hacker Assets Portal Team

Sagar Samtani
- 3rd Year Ph.D. and SFS Student
- University of Arizona

Kory Chinn
- Undergraduate Senior
- University of Arizona

Cathy Larson
- Former Associate Director, AI Lab
- University of Arizona

Dr. Hsinchun Chen
- Regents’ Professor
- Director, AI Lab
- University of Arizona
Introduction: “Know Your Enemy”

• Recent years have seen a significant increase in cybersecurity education initiatives.

• One novel way to enhance cybersecurity education and bolster future cyber-defenses is to directly study tools disseminated in online hacker communities.

• Online hacker forums allow hackers to share assets such as malicious tutorials, code, attachments.

• Spanning regions such as the US and Russia, there are tens of millions of posts in hundreds of forums made by millions of members.
  • Tens of thousands of malicious assets
Introduction – Hacker Asset Examples

Figure 1. Forum post with source code to create botnets

Figure 2. Forum post with BlackPOS malware attachment

Figure 3. Tutorial on how to create malicious documents
Introduction – AZSecure Hacker Assets Portal Objective

Given the rich nature of hacker forum data, we aim to design a web portal providing hacker forum contents and analysis for cybersecurity education, research, and training purposes.

• We achieve this goal by:
  • Identifying large English, Russian, and Arabic hacker forums
  • Extracting assets using advanced web crawling approaches
  • Analyzing assets using scalable text and data analytic methods
  • Developing a portal allowing users search, download, and analyze assets
AZSecure Hacker Assets Portal – Data Testbed

• We use a Tor routed web crawler to automatically collect one Arabic, two Russian, and two English forums known for containing malicious assets (Table 1).

• 15,576 code, 14,851 attachments, and 987 tutorials posted between 2/7/05-10/31/16.

• In addition to integrating other forums, we update our collection monthly to continually identify new and emerging assets.

<table>
<thead>
<tr>
<th>Forum</th>
<th>Language</th>
<th>Date Range</th>
<th># of Posts</th>
<th># of Members</th>
<th># of source code</th>
<th># of attachments</th>
<th># of tutorials</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenSC</td>
<td>English</td>
<td>02/07/2005-02/21/2016</td>
<td>124,993</td>
<td>6,796</td>
<td>2,590</td>
<td>2,349</td>
<td>628</td>
</tr>
<tr>
<td>Xeksec</td>
<td>Russian</td>
<td>07/07/2007- 9/15/2015</td>
<td>62,316</td>
<td>18,462</td>
<td>2,456</td>
<td>-</td>
<td>40</td>
</tr>
<tr>
<td>Ashiyane</td>
<td>Arabic</td>
<td>5/30/2003 – 9/24/2016</td>
<td>34,247</td>
<td>6,406</td>
<td>5,958</td>
<td>10,086</td>
<td>80</td>
</tr>
<tr>
<td>tuts4you</td>
<td>English</td>
<td>6/10/2006 – 10/31/2016</td>
<td>40,666</td>
<td>2,539</td>
<td>-</td>
<td>2,206</td>
<td>38</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td></td>
<td><strong>02/07/2005- 10/31/2016</strong></td>
<td><strong>590,699</strong></td>
<td><strong>47,492</strong></td>
<td><strong>15,576</strong></td>
<td><strong>14,851</strong></td>
<td><strong>987</strong></td>
</tr>
</tbody>
</table>

Table 1. Summary of AZSecure Hacker Assets Portal System Data
We use two automatic methods to sort assets (Figure 4).

First, we trained a Support Vector Machine (SVM) with 1,000 code files to classify hacker code into 10 languages:
- Java, Python, C/C++, HTML, Delphi, VB, SQL, Ruby, and Perl
- SVM outperformed other classifiers in standard metrics (Table 2)

We then use Latent Dirichlet Allocation (LDA) to each asset category to identify major themes (e.g., DDoS, Zeus, etc.).

Six SFS students evaluated accuracy of LDA results and reached a Cronbach’s alpha of 0.9393, indicating a high level of consistency.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>98.20</td>
<td>96.36</td>
<td>98.20</td>
<td>98.28</td>
</tr>
<tr>
<td>k-Nearest Neighbor</td>
<td>64.00</td>
<td>83.47</td>
<td>64.00</td>
<td>72.24</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>86.00</td>
<td>88.57</td>
<td>86.00</td>
<td>87.26</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>82.60</td>
<td>86.41</td>
<td>82.60</td>
<td>84.42</td>
</tr>
</tbody>
</table>

Table 2. Benchmark Classifier Evaluation Results
AZSecure Hacker Assets Portal System Design and Features

**Data Collection and Analytics**
- Latent Dirichlet Allocation (LDA) and Support Vector Machine (SVM) Analytics
- 987 tutorials, 15,576 source code, and 14,851 attachments

**Web Hosting and Access**
- AWS

**System Functionalities**
- Browsing
- Searching
- Downloading

**System Analytics**
- Cyber Threat Intelligence Dashboard
- VirusTotal Malware Analysis

Figure 5. AZSecure Hacker Assets Portal System Design and Features
Tutorial Data Collection Summary

<table>
<thead>
<tr>
<th>Tutorial Category</th>
<th>Count</th>
<th>Examples of Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Website Exploitation</td>
<td>348</td>
<td>SQL Injection, XSS attacks</td>
</tr>
<tr>
<td>System Exploitation</td>
<td>230</td>
<td>BIOS hacking, rootkit creation, shellcode, spoofing files</td>
</tr>
<tr>
<td>Carding</td>
<td>201</td>
<td>Carding, bank hacking</td>
</tr>
<tr>
<td>Network Exploitation</td>
<td>112</td>
<td>Nmap scanning, Wireshark, DDoS</td>
</tr>
<tr>
<td>Password Cracking</td>
<td>43</td>
<td>Brute forcing, password cracking approaches</td>
</tr>
<tr>
<td>Malware/Viruses</td>
<td>22</td>
<td>Malware analysis, detecting malware</td>
</tr>
<tr>
<td>Penetration Testing</td>
<td>13</td>
<td>Metasploit trainings, Google hacking</td>
</tr>
<tr>
<td>Mobile Exploitation</td>
<td>8</td>
<td>Android Malware</td>
</tr>
<tr>
<td>Cryptography</td>
<td>4</td>
<td>Basics of cryptography</td>
</tr>
<tr>
<td>Reverse Engineering</td>
<td>2</td>
<td>Basics of reverse engineering</td>
</tr>
<tr>
<td>Social Engineering</td>
<td>2</td>
<td>Social engineering psychology</td>
</tr>
<tr>
<td>Phishing</td>
<td>2</td>
<td>Basics of phishing</td>
</tr>
</tbody>
</table>

Table 2. Summary of Tutorial Data Collection Content

- Tutorials can provide the most direct cybersecurity education.

- We currently have 987 tutorials in 11 categories (Table 2).

- Tutorials teach various topics including:
  - Carding
  - SQL injections
  - Password cracking
  - Creating Android malware
  - Phishing
  - DDoS
## Code and Attachment Data Collection Summary

<table>
<thead>
<tr>
<th>Asset Type</th>
<th>Exploit Type</th>
<th>Count</th>
<th>Examples of Exploits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Code</td>
<td>System</td>
<td>9,746</td>
<td>Crypters, shellcode, DLL injections, Remote Administration Tools (RATs)</td>
</tr>
<tr>
<td></td>
<td>Website</td>
<td>5,598</td>
<td>Content management system (CMS) exploits, SQL Injections</td>
</tr>
<tr>
<td></td>
<td>Network</td>
<td>232</td>
<td>Bots/botnet/DDoS</td>
</tr>
<tr>
<td>Attachments</td>
<td>System</td>
<td>7,935</td>
<td>Zeus Malware, RATs, binders, Crypters, keyloggers</td>
</tr>
<tr>
<td></td>
<td>Website</td>
<td>3,112</td>
<td>Cross-site scripting (XSS), website backdoors, website defacing, phishing</td>
</tr>
<tr>
<td></td>
<td>Network</td>
<td>2,555</td>
<td>Bots/botnet/DDoS, firewall exploits, flooders</td>
</tr>
<tr>
<td></td>
<td>Database</td>
<td>1,039</td>
<td>SQL payloads, dumpers, SQLmap</td>
</tr>
<tr>
<td></td>
<td>Mobile</td>
<td>210</td>
<td>Android dumpers, crackers, malware, and pentests</td>
</tr>
</tbody>
</table>

- Students can use code and attachment assets to understand how tools are created, implemented, and operated.

- Code assets include:
  - Crypters
  - DLL injections
  - DDoS

- Attachments contain exploits such as:
  - Zeus
  - Android malware
  - Remote administration tools
  - Botnets
  - Keyloggers

Table 3. Summary of Source Code and Attachment Collection Content
Cyber Threat Intelligence (CTI) Applications

- For each asset, we detail which category of cyber-asset it targets (e.g., database, web, etc.) and where, when, who posted the asset in a CTI dashboard (Figure 6).
- Assuming an organization understands their own systems, hacker assets can create proactive CTI to inform future cyber-defenses.
- For example, an organization can improve mobile device security given the recent increase in mobile malware.
  - Can also identify key threat actors to monitor

Figure 6. (a) Selecting a specific time range for mobile malware, (b) list of mobile malware in selected range, and (c) key threat actors for selected malware.
Please access at: [http://www.azsecure-hap.com/](http://www.azsecure-hap.com/)
OR
Contact Sagar Samtani at sagars@email.arizona.edu

New users will need to enter their name, organization, position, and intended use to gain portal access.

We will then evaluate and confirm portal access.
AZSecure Hacker Underground Economy Collection and Analytics

Weifeng Li, Hsinchun Chen
Artificial Intelligence Lab, The University of Arizona
March 31st, 2017

*This work is supported by the National Science Foundation under Grant DUE-1303362 and SES-1314631
Hacker Underground Economy

- International online black markets for hacking services and tools
- Provides comprehensive support for conducting data breach crimes:
  - 2013: Target;
  - 2014: Home Depot, Chase;
  - 2015: Anthem;
  - 2016: Yahoo
- Common platforms: hacker forums, DarkNet marketplaces, carding shops

Greater Underground Assets Supply

Lower Cybercrime Barrier

Underground Economy

Greater Underground Assets Demand

More Cybercriminals
What is in the underground economy?

POS Skimmer

Target POS device: Verifone vx510/vx670

YouTube Tutorials

Method of Payment: Liberty Reserve

ATM Skimmer

Accessories

Tutorials

Blank Credit/Debit Cards (Plastics)

Features

EMV encoder

Sold in batch

EMV encoder

sendspace.com/f

Good luck !!!
## Collection Summary

### Forums

<table>
<thead>
<tr>
<th># Forums</th>
<th># Products</th>
<th># Users</th>
<th># Topics</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>169,009</td>
<td>257,183</td>
<td>414,530</td>
<td>English/Russian/Arabic</td>
</tr>
</tbody>
</table>

### DarkNet Marketplaces

<table>
<thead>
<tr>
<th># Markets</th>
<th># Products</th>
<th># Users</th>
<th># Reviews</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>80,590</td>
<td>5,528</td>
<td>690,411</td>
<td>English/Russian/Dutch</td>
</tr>
</tbody>
</table>

### Carding Shops

<table>
<thead>
<tr>
<th># Shops</th>
<th># Listings (cards, SSNs)</th>
<th># Users</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>1,401,708</td>
<td>N/A</td>
<td>English/Russian</td>
</tr>
</tbody>
</table>

### In Total:

<table>
<thead>
<tr>
<th># Platforms</th>
<th># Products</th>
<th># Participants</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>1,651,307</td>
<td>262,711</td>
<td>English/Russian/Arabic/Dutch</td>
</tr>
</tbody>
</table>

* Hacker assets include but are not limited to: malware (encrypter/ransomware, Trojan, exploit), zero-day vulnerabilities, POS/ATM skimmer, stolen credit/debit card, fake documents (driver’s license, SSN), etc.
Sample Stolen Data in Collection

- Dump: magnetic strip information (for fraud purchase in store)
  
  | Card Number | 5177220046948089^FARAH TUKAN (Track 1) |
  | Expiration | ^150410110000000000000000149 (Track 2) |
  | 5177220046948089=1504101100000149 (Track 3) |

- CC/CVV: magnetic strip information
  
  Card Number: 4266841209090735
  Expire Date: 01/2012
  CVV: 131
  Cardholder Name: Walter Leger
  Address: 4701 Rue Laurent
  City: Metairie – LA – 70002

- Fullz: magnetic strip information
  
  | Driver’s License | James|Gayner|28540 Doyle Creek Rd.|Saint Marys|KS|66536|785-437-2803|362-82-4079|k00073521|KS|03-17-1967|JPGayner@yahoo.com|1ps72bn93d |

- Health insurance records
  
  PRIMARY, Cigna Healthcare,U04197556,2461898,
  SECONDARY, UGA Athletic Dept.,254718352,,650 West Conway Dr.,27y,4/8/1989,,Atlanta,,MARGARET,A,MCWHIRTER,,(404)401-3108,,F,254-71-8352,GA,30327

Ramification:
- Fraudulent purchase in store
- Fraudulent purchase online
- Fraudulent loan application/tax return
- Fraudulent healthcare claim
## Summary of Major Data Breach Services in Collection

<table>
<thead>
<tr>
<th>Category</th>
<th>Service</th>
<th>Examples</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Infrastructure</strong></td>
<td>Malware</td>
<td>POS malware; ATM skimmers</td>
<td>$300~5000</td>
</tr>
<tr>
<td></td>
<td>Phishing</td>
<td>Phishing emails; scam sites</td>
<td>$2.5~100/wk</td>
</tr>
<tr>
<td></td>
<td>Botnets</td>
<td>Hosting relays for stolen cards</td>
<td>$2~60/hr</td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>Payment Cards</td>
<td>Dumps; CC/CVV</td>
<td>$0.1~25</td>
</tr>
<tr>
<td></td>
<td>Identities (Fullz)</td>
<td>Social Security Numbers; driver's license; insurance cards</td>
<td>$1~260</td>
</tr>
<tr>
<td></td>
<td>Credentials</td>
<td>Bank accounts; Paypal accounts</td>
<td>$1~300</td>
</tr>
<tr>
<td><strong>Cashing</strong></td>
<td>Forging</td>
<td>Blank credit cards; driver’s license template</td>
<td>$40~110</td>
</tr>
<tr>
<td></td>
<td>Change of Billing (COB)</td>
<td>Change of billing address for carders to make purchases</td>
<td>$35~140</td>
</tr>
<tr>
<td></td>
<td>Drop</td>
<td>Location carders can have illicitly purchased goods sent to</td>
<td>~50% Royalty</td>
</tr>
</tbody>
</table>

*Table 1. Common Hacking Services and Their Prices*
Analytics: Key Seller Identification

Figure 1. The AZSecure Key Seller Identification Framework
Figure 2. Thread Classification Performance

(a) Malware Advertisements

(b) Stolen Data Advertisements

Figure 3. Seller Rating Performance

(a) Positive Sentiment

(b) Negative Sentiment
### Who are the key sellers?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Seller</th>
<th>Score</th>
<th>User</th>
<th>Score</th>
<th>Seller</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Antichat</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>LEOnidUKG</td>
<td>5</td>
<td>inferno[DGT]</td>
<td>3.6</td>
<td>@NoFrag@</td>
<td>1.8</td>
</tr>
<tr>
<td>2</td>
<td>VISMU</td>
<td>4.5</td>
<td>amorales</td>
<td>3.5</td>
<td>k0lbasa</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>gogol</td>
<td>4</td>
<td>«DEXTER»</td>
<td>3.4</td>
<td>Doktor_radosti</td>
<td>2</td>
</tr>
<tr>
<td><strong>CrdPro</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>HackingAll</td>
<td>4</td>
<td>Rescator</td>
<td>4.4</td>
<td>N1K70</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>balt</td>
<td>4</td>
<td>Faaxxx</td>
<td>4</td>
<td>1vanu4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>SunSeller</td>
<td>4</td>
<td>ResellerInc</td>
<td>4</td>
<td>MID</td>
<td>2</td>
</tr>
<tr>
<td><strong>Zloy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>PerfectCrypt</td>
<td>5</td>
<td>BigBuyer</td>
<td>4</td>
<td>root</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>DiXakMan</td>
<td>4</td>
<td>Buyers11</td>
<td>4</td>
<td>w370w370</td>
<td>1.6</td>
</tr>
<tr>
<td>3</td>
<td>DjVellf</td>
<td>4</td>
<td>sellcc</td>
<td>4</td>
<td>gorpen</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2. Top 3 Best/Worst Malware and Stolen Data Sellers for Each Forum
• Before the announcements of data breach events, bulks of breached data already appear in key sellers’ shops for sale.
How much is each card worth?

<table>
<thead>
<tr>
<th>Card Features: Brand, Type, Mark, Bank</th>
<th>Price Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visa Electron Card</td>
<td>$21.06</td>
</tr>
<tr>
<td>American Express Card</td>
<td>$19.41</td>
</tr>
<tr>
<td>World Elite MasterCard for Business</td>
<td>$17.11</td>
</tr>
<tr>
<td>Corporate Purchasing Card</td>
<td>$13.15</td>
</tr>
<tr>
<td><strong>Base Price</strong></td>
<td><strong>$7.48</strong></td>
</tr>
<tr>
<td>Bank of America Card</td>
<td>$2.63</td>
</tr>
<tr>
<td>Debit Card</td>
<td>$2.00</td>
</tr>
<tr>
<td>Credit Card</td>
<td>$1.72</td>
</tr>
<tr>
<td>J.P. Morgan Chase Card</td>
<td>–$1.78</td>
</tr>
<tr>
<td>Standard Card</td>
<td>–$2.43</td>
</tr>
<tr>
<td>Prepaid Card</td>
<td>–$4.72</td>
</tr>
<tr>
<td>Visa/MasterCard Classic Card</td>
<td>–$4.77</td>
</tr>
</tbody>
</table>

Table 3. Selected Card Features Affecting Card Price in the Underground Economy

* Results were obtained using standard linear regression model with significance level of 0.001.