ISI Research and Analysis Tool Inventory

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. This document provides a summary of tools that can aid researchers in performing data driven security analytics. The presented tools are not exhaustive of all tools that currently exist in the data analytics landscape. Rather, they reflect the tools used in the University of Arizona's Artificial Intelligence Lab's past security informatics research. We organize the tools into three major sections based on a traditional data analytics pipeline: (1) collection and storage tools; (2) pre-processing and analytics tools; and (3) visualization tools. For each category, we provide a short summary of the typical types of tasks that are completed in that phase of the data analytics procedure followed by an inventory of tools that fall into that category. We provide the name of the tool, a link of where to download and get documentation for each tool. Note that researchers can select one of the tools with similar functionalities based on personal preference (such as WEKA vs. RapidMiner). We also select a set of ISI papers which have used the listed tools.

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.) and index and store it in a database or some other format which can be can be retrieved and used for pre-processing and further analytics. This section details some of the packages and tools that can be used to collect and store data. On a high level, the collection process comprises three steps to pull from the online sources into the database: extract, transform, and load (ETL).

Collection Process: Extract

The first part of the collection process involves extracting the data from the online source. Depending on the source system, different techniques can be used for extraction. Some sources may also have anti-crawling measures built in. We provide several techniques and strategies to counter some of these measures.

Spidering tools

Tool	**Details**
Offline Explorer	Offline Explorer (OE) Pro is a useful tool we use for collecting forum and other web contents. OE provides a very useful GUI for creating and scheduling various crawling projects, built-in support for completing HTML login forms, and even supports routing traffic through proxy servers and the Tor network. Link: https://www.metaproducts.com/mp/offline_explorer.htm

cURL	cURL is a tool to transfer data from or to a server, using one of the supported protocols. cURL offers a busload of useful tricks like proxy support, user authentication, FTP upload, HTTP post, SSL connections, cookies, file transfer resume, Metalink, and more. Link: https://curl.haxx.se/
Wget	Wget is a free utility for non-interactive download of files from the Web. It supports HTTP, HTTPS, and FTP protocols, as well as retrieval through HTTP proxies. Its features include recursive download, conversion of links for offline viewing of local HTML, and support for proxies. Link: https://www.gnu.org/software/wget/

Packages for Customized Spiders

Package	**Programming Language**	**Details**
HtmlUnit	Java	HtmlUnit is a headless web browser written in Java. It allows high-level manipulation of websites written in Java code, including filling and submitting forms and clicking hyperlinks. It also provides access to the structure and the details within received web pages. HtmlUnit emulates parts of browser behaviour including the lower-level aspects of TCP/IP and HTTP. This headless browser can deal with HTTPS security, basic HTTP authentication, automatic page redirection and other HTTP headers. Link: http://htmlunit.sourceforge.net/
Selenium	Python	Selenium is a browser automation library. Selenium may be used for any task that requires automating interaction with the browser. Selenium makes direct calls to the browser using each browser’s native support for automation. Link: http://selenium-python.readthedocs.io/

Counter Anti-crawling Techniques

Anti-crawling Measure	**Description**	**Counter-measure**
User-agent Check	Shops verify the HTTP request comes from a legitimate user-agent (browser.)	Use packages that mimics the behavior of mainstream browsers.

User/password Authentication	Shops requires users to register and login before accessing the data. CAPTCHA is widely used to verify the user inputting the credential is a human-being.	Login the shop first and extract the corresponding cookies. With these cookies carried with HTTP request, we can bypass the login process.
Session Timeout	Shops automatically logout users that have been in the shop for too long.	Need human involvement to acquire and deploy renewed cookies.
IP Check	CloudFlare verifies the HTTP request comes from a legitimate IP address rather than a public known proxy, such as Tor.	Setup a private, dedicated proxy server to reroute our connections. The proxy server can be deployed in Digital Ocean as it is easily to deploy a new IP after the first IP is banned.
DDoS Prevention	CloudFlare detects possible DDoS signs and bans the suspicious IP address.	Set intervals between two successive requests; allow the private proxy server to change IP addresses easily

Collection Process: Transform

The second part of the collection process involves transforming the raw data into target data elements. These tools help parse the target data elements from the raw collected data, especially web pages.

Tool	**Details**
Regex	A regular expression, regex or regexpx is a sequence of characters that define a search pattern. Usually this pattern is then used by string searching algorithms for "find" or "find and replace" operations on strings. Link: http://www.regular-expressions.info/
JSoup	JSoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods. Link: https://jsoup.org/
BeautifulSoup	Beautiful Soup is a Python package for parsing HTML and XML documents (including having malformed markup, i.e. non-closed tags, so named after tag soup). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping. Link: https://www.crummy.com/software/BeautifulSoup/

urllib	This module provides a high-level interface for fetching data across the World Wide Web. In particular, the urlopen() function is similar to the built-in function open(), but accepts Universal Resource Locators (URLs) instead of filenames. Some restrictions apply — it can only open URLs for reading, and no seek operations are available. Link: https://docs.python.org/2/library/httplib.html

Collection Process: Load

The last part of the collection process involves loading the data into the data warehouse. Here are a list of common data warehouse implementations and their associated documentation.

Implementation	**Details**
MySQL	MySQL is an open-source relational database management system (RDBMS). Link: https://www.mysql.com/
MS SQL Server	Microsoft SQL Server is a relational database management system developed by Microsoft. Link: https://www.microsoft.com/en-us/sql-server/sql-server-2016
Oracle Database	Oracle Database is an object-relational database management system produced and marketed by Oracle Corporation. Link: https://www.oracle.com/database/index.html
Apache HBase	Apache HBase is an open-source, distributed, versioned, non-relational database modeled after Google's Bigtable. Apache HBase provides Bigtable-like capabilities on top of Hadoop. Use Apache HBase when you need random, real time read/write access to your Big Data. Link: https://hbase.apache.org/
Apache Hive	Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. Link: https://hive.apache.org/
MongoDB	MongoDB (from humongous) is a free and open-source cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemas. Link: https://www.mongodb.com/

Apache Lucene	Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform. Link: http://lucene.apache.org/core/

Pre-Processing and Analytics Tools

Before data can be analyzed, it often has to be pre-processed and transformed into a format which is conducive for analysis. Such a process often consumes the majority (70-75%) of the time in data analytic projects. Pre-processing tasks include, but are not limited to, cleaning, normalizing, transforming, tokenizing, extracting features, tagging parts of speech, etc. While custom scripts are often required in pre-processing, there are some general purpose tools that can help convert data into usable formats for analytics.

Once data has been pre-processed and converted into a format appropriate for analysis, the third phase in the data analytics pipeline focuses on analyzing the data to derive useful and interesting insights. Past security analytics research has employed dozens of analytical techniques ranging from simple summary statistics to complex algorithms such as deep learning. This results in a large range of tools that can be applied for security analytics. Many common data mining algorithms (e.g., SVM, Naive Bayes, k-means, regression, etc.) and general text mining applications (Named Entity Recognition, coreference resolution, etc.) are bundled into single packages such as WEKA or Natural Language Toolkit. However, there are various analytical approaches (e.g., hidden markov models, conditional random fields, social network analysis, etc.) that are not currently part of any general toolset, but part of a more specialized package. Those tools are also listed.

Tool Type	Tool Name	Programming Language	URL for Documentation	Notes
General Data Mining	WEKA	Java, GUI	http://www.cs.waikato.ac.nz/ml/weka/	One-stop tools that cover popular pre-processing, classification, and clustering algorithms. RapidMiner and WEKA can be used independently without a specific programming language.
	Scikit-Learn	Python	http://scikit-learn.org/stable/	
	RapidMiner	GUI	https://rapidminer.com/	
	R	R	https://www.r-project.org/	A widely used programming language and software environment for

				statistical computing and graphics. Various data pre-processing and analytics tools are supported by packages.
General Text Mining	Natural Language Toolkit (NLTK)	Python	http://www.nltk.org/	One-stop tools that cover word/sentence tokenization, POS tagging, parsing, chunking, named entity recognition, etc. NLTK has interfaces to call Stanford NLP tools.
	Stanford CoreNLP	Java	http://nlp.stanford.edu/software/	
	Apache OpenNLP	Java	https://opennlp.apache.org/	
Sentiment Analysis	SentiStrength	Java	http://sentistrength.wlv.ac.uk/	Estimates the strength of positive and negative sentiment in short texts.
Ontologies	WordNet	-	https://wordnet.princeton.edu/	English lexical database grouped into synonyms.
	SentiWordNet	-	http://sentiwordnet.isti.cnr.it/	Tagged WordNet with positivity, negativity, and neutrality for opinion mining.
Hidden Markov Models (HMM)	hmmlearn	Python	https://hmmlearn.readthedocs.io/en/latest/	General HMM package
	NLTK	Python	http://www.nltk.org/	Specialized in POS tagging
Conditional Random Fields (CRF)	Stanford NLP Group	Java	http://nlp.stanford.edu/software/	Stanford NER CRF has a CRF implementation
	CRF++	C++	https://taku910.github.io/crfpp/	General CRF package
	NLTK	Python	http://www.nltk.org/	Specialized in POS tagging, referring to a package pycrfsuite
Latent Dirichlet Allocation	Mallet	Java	http://mallet.cs.umass.edu/mallet-tutorial.pdf	Command line based tool that can perform standard LDA

(LDA)	Stanford Topic Modelling Toolbox	GUI	http://nlp.stanford.edu/software/tmt/tmt-0.4/	GUI based tool that supports LDA, labelled LDA, partially labelled LDA, and calculating perplexity. Can also perform temporal LDA
	Gensim	Python	https://radimrehurek.com/gensim/tutorial.html	Allows users to perform Latent Semantic Analysis and LDA using Python. Useful when integrating LDA with other applications in Python
Social Network Analysis	UCINET	GUI	https://sites.google.com/site/ucinetsoftware/home	Licensed software (minimum $40) that can handle medium sized networks (2 millions nodes max)
	Gephi	GUI	https://gephi.org/	Open source GUI based software that can handle larger data sizes than UCINET. Can read directly from databases
	NetworkX	Python	https://networkx.readthedocs.io/en/stable/	Python based network analysis tools. Can read from a variety of data sources. Allows for significant customization compared to other tools
Word2vec	Gensim	Python, C	http://radimrehurek.com/gensim/models/word2vec.html	Word2vec is a two-layer neural net that processes text. Its input is a text corpus and its output is a set of vectors: feature vectors for words in that corpus. While Word2vec is not a deep neural network, it turns text into a numerical form that deep nets can understand.
	DL4J	Java, Scala	https://deeplearning4j.org/word2vec.html	
Deep Learning	Keras	Python	https://keras.io/	High-level neural networks library running on top of either TensorFlow or Theano. Recommended for fast experimentation.

	TensorFlow	Python, C++	https://www.tensorflow.org/tutorials/	Low-level implementation for deep learning models
	Theano	Python	http://deeplearning.net/software/theano/	Low-level implementation for deep learning models

Visualization Tools

The final stage in the data often incorporates a visualization component, where researchers will utilize various tools to create diagrams. Desktop software provide turnkey solutions to manage, connect, pivot data and render predefined types of visualizations in the GUI. For better customizability, lightweight toolkits, packages, and online services can be implemented along with analytical scripts.

Desktop Visualization Software

Tool	**Cost**	**Descriptions**	**Tutorials**
Microsoft Excel	License Required	Excel supports charts, graphs, generated from specified groups of cells. Excel 2010 and later support Pivot Table, which enables geo-map plotting as well as interactive visualizations.	https://support.office.com/en-us/article/Power-View-Explore-visualize-and-present-your-data-98268d31-97e2-42aa-a52b-a68cf460472e
Tableau	Free Education License	Tableau queries relational databases, cubes, cloud databases, and spreadsheets and then generates a number of graph types that can be combined into dashboards and shared over a computer network or the internet.	https://www.tableau.com/learn/training
ParaView	Free, Open-source	Users can quickly build visualizations to analyze their data using qualitative and quantitative techniques. The data exploration can be done interactively in 3D or programmatically using its batch processing capabilities. ParaView was developed to analyze extremely large datasets using distributed memory computing resources.	http://www.paraview.org/tutorials

Lightweight Toolkits, Packets, and Online Services

Tool Type	**Tool Name**	**Programming Language**	**URL for Documentation**
General Data Visualization Toolkits	General data visualization toolkits enabled users to customize their visualization components (e.g., point, line, axes, legends, data layout, color coding) programmatically. Matplotlib, Seaborn, pandas, ggplot2 provide basic visualization templates (e.g., scatterplot, bar chart) for fast visualization implementation.		
	Visualization Toolkit (VTK)	C++, Python, Java	http://www.vtk.org/
	OpenFrameworks (OF)	C++	http://openframeworks.cc/ofBook/chapters/foreword.html
	Processing	Java, Python, Javascript	https://processing.org/tutorials/
	Matplotlib	Python	http://matplotlib.org/index.html
	Seaborn	Python	http://seaborn.pydata.org/
	pandas	Python	http://pandas.pydata.org/
	ggplot2	R	http://ggplot2.org/
Word Cloud	Word cloud is a graphical representation of word frequencies. It can be used to visualize most frequently used keywords in the corpus.		
	Wordle	Online, Javascript	http://www.wordle.net/
Geo-Map Tools	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.		
	Mapbox	Online, Javascript	https://www.mapbox.com/
	geoplotlib	Python	https://github.com/andrea-cuttone/geoplotlib
	choroplethr	R	https://github.com/trulia/choroplethr

Network Visualization Tools	Network visualization tools can visualize the relationship between data attributes or different data sources. The built in layout algorithms automatically generate visually pleasing graphs.		
	Gephi	GUI, Java	https://gephi.org
	networkx	Python	https://networkx.github.io/
	graph-tool	Python	https://graph-tool.skewed.de/
	igraph	R	http://igraph.org/
Front-end Visualization Tools	These tools provides solutions to embed static/interactive visualization on the webpage. Predefined templates are available so they are light-weight design tools compared with general visualization toolkits.		
	D3.js	Javascript	http://alignedleft.com/tutorials/d3
	Google Chart	Javascript	https://developers.google.com/chart/
		R (googleVis)	https://cran.r-project.org/web/packages/googleVis/vignettes/googleVis_examples.html
	Datawrapper	Online, Javascript	https://datawrapper.de/
	Infogram	Online, Javascript	https://infogr.am/
	Plotly	Online, Javascript, R, Python	https://plot.ly/
Interactive Visualization Tools	Interactive visualization tools support user interactions such as highlighting, zooming, and panning. Interaction visualization is a good way to present data with different granularities of details or with time-series changes.		
	Bokeh	Python	http://bokeh.pydata.org/en/latest/docs/user_guide.html#userguide
	ggvis	R	http://ggvis.rstudio.com/
	visNetwork	R	http://datastorm-open.github.io/visNetwork/

Color Selection (Aesthetic)	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.		
	Color Brewer 2	Online	http://colorbrewer2.org
	Palettable	Python	https://jiffyclub.github.io/palettable/
	RColorBrewer	R	https://cran.r-project.org/web/packages/RColorBrewer/index.html

Example ISI Papers

To show the research context of applying the listed tools, we reviewed research papers from 2016 and 2015 IEEE ISI (56 and 47 papers respectively), 2016 FOSINT-SI (8 papers), and 2015 ISI-ICDM (10 papers). Following the structure of this document, tools are categorized into collection, storage, pre-processing, analytics, and visualization tools. We selected representative papers to show how those tools can be used together to support research. Note that around 70 percent of the papers we reviewed did not specify the tools they used, especially for storage and visualization, or only mentioned the techniques instead of the tools for implementation.

Paper	**Collection and Storage**	**Pre-Processing and Analytics**	**Visualization**
Samtani et al. (2016)	Offline Explorer, MySQL, Regex	RapidMiner, Stanford Topic Modelling Toolbox	Tableau, D3.js
Grisham et al. (2016)	Selenium, MySQL	Stanford Topic Modelling Toolbox	-
Benjamin & Chen (2016)	Offline Explorer, MySQL, Regex	Word2vec	-
Benjamin & Chen (2014)	IRC Bots	WEKA	-
Samtani & Chen (2016)	Offline Explorer, MySQL, Regex	Gephi	Gephi
Solaimani et al. (2016)	MongoDB	CoreNLP, WordNet	-
Dobolyi & Abbasi (2016)	PhishTank API, Wget	R	R

Andrew Park et al. (2016)	SQLite	Apache OpenNLP, SentiStrength	-

References

1. Samtani, S., & Chen, H. (2016, September). Using social network analysis to identify key hackers for keylogging tools in hacker forums. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 319-321). IEEE.
2. Grisham, J., Barreras, C., Afarin, C., Patton, M., & Chen, H. (2016, September). Identifying top listers in Alphabay using Latent Dirichlet Allocation. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 219-219). IEEE.
3. Samtani, S., & Chen, H. (2016, September). Using social network analysis to identify key hackers for keylogging tools in hacker forums. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 319-321). IEEE.
4. Samtani, S., Chinn, K., Larson, C., & Chen, H. (2016, September). AZSecure Hacker Assets Portal: Cyber threat intelligence and malware analysis. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 19-24). IEEE.
5. Benjamin, V., & Chen, H. (2016, September). Identifying language groups within multilingual cybercriminal forums. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 205-207). IEEE.
6. Dobolyi, D. G., & Abbasi, A. (2016, September). PhishMonger: A free and open source public archive of real-world phishing websites. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 31-36). IEEE.
7. Solaimani, M., Salam, S., Mustafa, A. M., Khan, L., Brandt, P. T., & Thuraisingham, B. (2016, September). Near real-time atrocity event coding. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 139-144). IEEE.
8. Park, A. J., Beck, B., Fletche, D., Lam, P., & Tsang, H. H. (2016, August). Temporal analysis of radical dark web forum users. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on* (pp. 880-883). IEEE.