AZSecure-data.org

Summer Data Release

The AZSecure-data.org data portal has published new intelligence and security informatics-related data sets. Security researchers are encouraged to explore the collections and announce their availability to their students and other researchers.

With funding from the National Science Foundation, this Data Infrastructure Building Blocks for Intelligence and Security Informatics (“DIBBs for ISI”) project is developing a research infrastructure for use by ISI scientists, students, and others studying a wide range of cyber- and security-related issues, whether in the computer, information, or social sciences. The project portal, AZSecure-data, provides researchers and students with free access to a wide variety of relevant data sets.

The newly added data sets include the following:

- A collection of phishing attack data organized by prominent targeted brands joins our growing collection of phishing data. This data set contains time series website data from 2006 through 2015 for 178 prominent targeted brands, with URL and Whois information for each phishing attack. The data includes nearly 1.5 million attack URLs for the study of phishing and other forms on internet fraud.
- Recent hacking forums in both Russian and English cover hacking and carding topics from September 2007 through September 2015 with a combined total of 138,100 forum posts. Hacking forums can support the study of hacker behavior and how hackers learn from each other, the formation of social networks, relationships with the underground economy, and more.
- Two of the newest collections focus on the Chinese underground market – Baidu Forums data is presented in English, containing over 53,000 posts from January 2006 to March 2016, while QQ Chat Logs are presented in Chinese and cover March 18, 2006 through April 4, 2016. Baidu and QQ each provide a general picture of the Chinese underground economy. Baidu provides details of products and services for sale and their customer reviews; QQ provides logs of cybercrimal communications in hacker groups. These data sets, consisting of forum postings and chat logs, can be used to study cyber security, social issues, authorship analysis, multi-lingual text analytics, and a wide array of other topics.
- A collection of 207 patriot, militia, hate and linked websites, from 74 from groups identified by the Southern Poverty Law Center in 2009 as promoting extreme social perspectives, and an additional 133 linked websites, collected by the University of Arizona Artificial Intelligence Lab for a study examining the relationships between virtual and real organizations. The websites can also be used to study rhetoric and communication, group dynamics, extreme social movements, and other topics, in information and the social sciences.
- Eighteen additional forum datasets incorporated into our existing GeoWeb collection add over 3.5M posts from nearly 120,000 users between the dates of January 2001 and April 2012, bringing the total forums in that collection up to 65. GeoWeb datasets, consisting of forum postings, can be used, for example, to study and analyze state as well as non-state social movements, current opinions, and social communications and their dynamics in at-risk countries and regions.
- CTU-13 and ISOT Botnet are the first datasets on the portal containing network traffic captures (botnet, normal, and background traffic) from a variety of sources collected from October 2004 through January 2005 and in August 2011. These data sets can be used to investigate malware and other malicious traffic.

These new collections join existing data sets comprising over one thousand phishing and spoof websites for classification tasks and fraud detection and prediction, and the earlier Dark Web and Geo Web collections of web forums comprising tens of millions of forum postings from multiple countries in many languages, for classification, network analysis,

Hsinchun Chen, Ph.D.
Regents' Professor, Thomas R. Brown Chair of Management and Technology, and Director of the Artificial Intelligence Lab, University of Arizona.
PI and Director for Data Infrastructure Building Blocks for Intelligence and Security Informatics (DIBBs for ISI).
hchen@eller.arizona.edu
University of Arizona
Tucson, AZ 85721 USA
520.621.2748 tel
520.521.2433 fax
ai.arizona.edu

Interested in learning more?
Join our Intelligence and Security Informatics Group on LinkedIn

Do you have data to share?
We can help! Send an inquiry to ailab@email.arizona.edu.

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

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

sentiment and affect analysis, and more. (See more about the Dark Web and Geo Web projects on the AI Lab website.) These newest additions increase both the depth and the breadth of the data available and greatly enrich the amount of data available for security researchers.

AZSecure-data can help you fulfill sponsor requirements to share data. If you would like to join the DIBBs for ISI Honor Roll and donate a data set, contact us at ailab@email.arizona.edu. We make it as easy as possible to deposit your data!

Not interested in receiving further news releases? Send an email from your account to ailab@email.arizona.edu, with a subject line "unsubscribe azsecure-data" and include your email address.