SaTC: CORE: Small: Cybersecurity Big Data Research for Hacker Community: A Topic and Language Modeling Approach

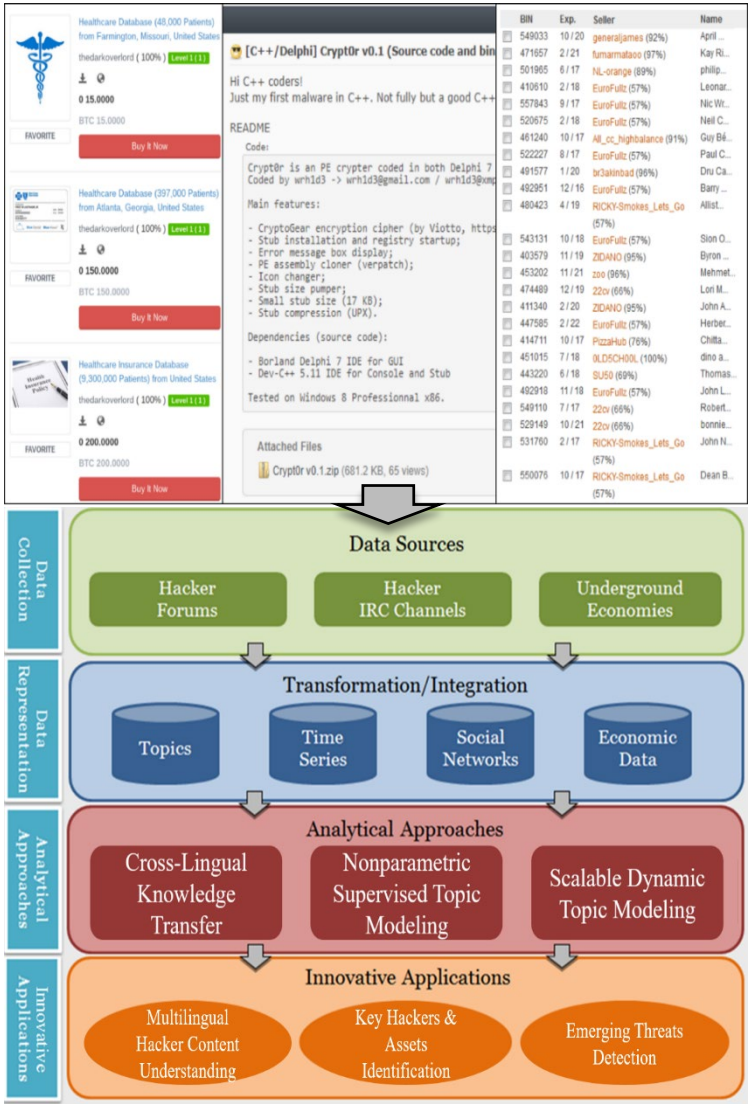
Challenge:

- Conventional Cyber Threat Intelligence (CTI) from historical attack data lacks proactive intelligence and ignores threat actors.
- Hacker communities make excellent candidates for research due to the salient information about the cybercriminal assets (e.g., malware, tutorials) shared by hackers, but presents challenges:

1. Technical difficulties in hacker community collection: anti-crawling measures
2. Heterogeneity and covert nature of the data elements and their subtle linkages
3. Subcultural nature of terms and concepts embedded in the hacking community across multiple foreign languages

Solution:

- A large and comprehensive collection of significant international online hacker communities
- An advanced data collection method counteracting anti-crawling measures such as CAPTCHA
- An adversarial deep representation approach to learning the language-invariant representations from content in English and non-English hacker communities
- A nonparametric supervised topic modeling method for examining customer reviews of hacker assets
- A scalable dynamic topic modeling technique designed for incorporating expert knowledge of hacker communities



Scientific Impact:

- Methodological contributions in hacker content analysis to assist practitioners developing proactive CTI from critical hacker communities
- A rich, large-scale, longitudinal, open-source collection of diverse hacker community field data to support data-driven research
- Knowledge advancement in deep transfer learning, deep generative modeling, supervised topic modeling, dynamic topic modeling, neural variational inference
- Applicability to other big data problems, especially in the social media context

Broader Impact and Broader Participation:

- Research dissemination to cybersecurity communities: Intelligence and Security Informatics (ISI), NSF Scholarship-for-Service (SFS), National Cyber-Forensics & Training Alliance (NCFTA), and The Society for the Policing of Cyberspace (POLCYB).
- Student training and professional development: AZSecure SFS, NSA designated Center of Academic Education in Cyber Defense (CAE-CD) courses at UA, UA's MS in Cybersecurity program, UGA's area of emphasis in Information Security

NSF Award#1936370
PI: Dr. Hsinchun Chen (hsinchun@email.arizona.edu)
Co-PI: Dr. Weifeng Li (weifeng.li@uga.edu)