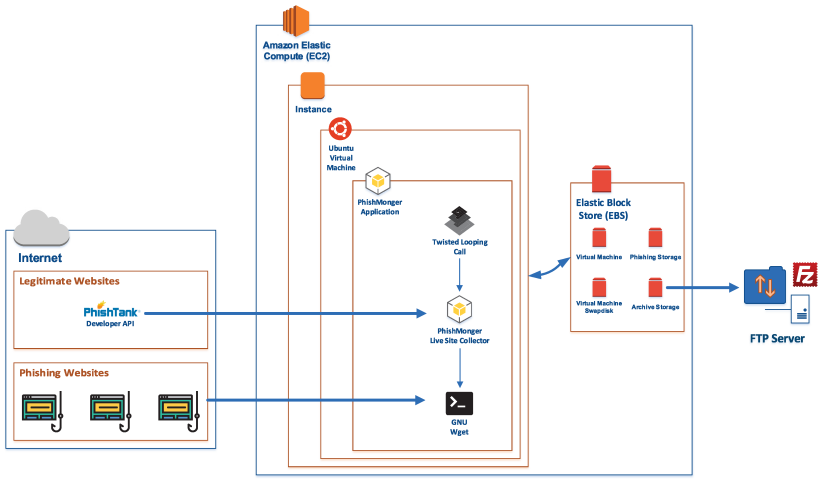
PhishMonger: A Free and Open Source Public Archive of Real-World Phishing Websites

David G. Dobolyi & Ahmed Abbasi
Center for Business Analytics, McIntire School of Commerce, University of Virginia

Introduction

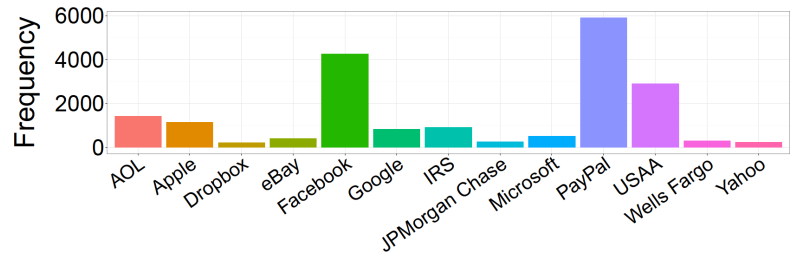
The number of active, online phishing websites continues to grow unabated in recent years. This has created an ever-increasing security risk for both individual and enterprise users in terms of identity theft, malware, financial loss, etc. Although resources exist for tracking, cataloguing, and blacklisting these types of sites (e.g., PhishTank.com), the ephemeral nature of phishing websites makes in-depth analysis exceptionally difficult. In order to better understand how these phishing sites exploit user and system weaknesses, we have crafted a platform named ***PhishMonger*** to capture live phishing websites in real-time on an ever-present, rolling basis. As of run **3649**, our growing corpus of verified phishing websites currently encompasses over **171,360 sites** involving **129 targeted brands**, which span **19,690,341 files and folders** and utilize **200GB** of compressed storage. The corpus is freely available online at **http://www.azsecure-data.org/phishing-websites.html**

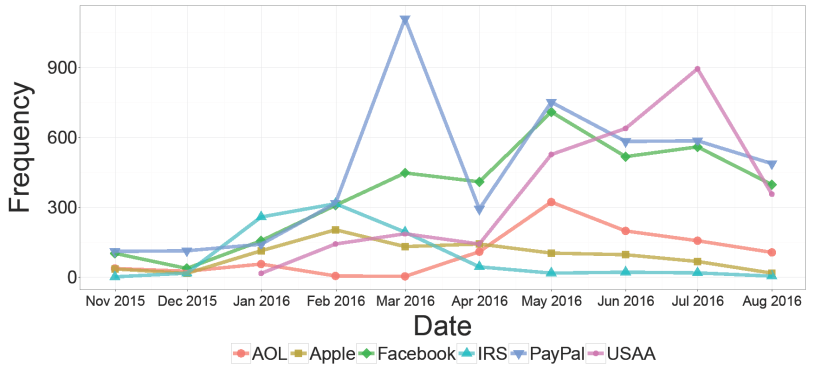
System Diagram



Corpus Details

File Extension	Description	Category	n
png	Portable Network Graphics	Graphics	3,169,772
html	HyperText Markup Language	Text	1,304,574
jpg	Joint Photographic Experts Group	Graphics	1,251,227
gif	Graphics Interchange Format	Graphics	1,208,424
js	JavaScript Code	Text	947,420
css	Cascading Style Sheet	Text	776,404
ttf	True Type Font	Font	210,197
svg	Scalable Vector Graphics	Graphics	176,856
ico	Icon	Graphics	139,564
woff	Web Open Font Format	Font	134,308





Acknowledgements: This research has been supported in part by the following U.S. National Science Foundation grant: ACI-1443019 "DIBBs for Intelligence and Security Informatics Research Community." Additionally, we would like to thank our Data Infrastructure Buildings Blocks (DIBBs) partner institutions including the University of Arizona's Artificial Intelligence Lab, which serves as the overall project lead (https://ai.arizona.edu/research/dibbs#portal), as well as collaborators at Drexel University, the University of Texas-Dallas, and the University of Utah.