Malucrawl

A Trending Topic Malware Crawler

Week 4 Progress Seminar

Introduction

- Trending Topics
  - Twitter
  - Media
  - Google
- Is Malware targeted towards trends?
  - Emma Watson => Malware?
  - McAfee Report
- Build Distributed Framework

Literature Review: deSEO

- First systematic study of search-result poisoning attacks and detection.

- SEO: Search Engine Optimization techniques
  - Optimization of page rank / Search-result poisoning
  - Cloaking techniques from attacker
  - Classified into White-hat and Black-hat

- Components of an SEO attack:
  - Automated generation of “relevant” content
  - Targeting multiple trending keywords
  - Creating dense link structures to increase page rank

Literature Review: deSEO

- deSEO: automatically detect SEO attacks and campaigns
  - History-based detection
  - Clustering of suspicious domains
  - Group analysis

- Result:
  - More than 300 billion URLs dataset
  - Result: 957 unique compromised domains, 15,482 malicious URLs.

Lit. Review: Automatic Malware Collecting System

- A system collects search keywords and the malicious code which uses the keywords.
- Malicious code collection approaches:
  - Passive method
  - Active method: Low interaction & High interaction
- The automatic malicious code collecting system:
  - Active Hybrid Method
  - Collect search keywords
  - Filter search results by reviewing URL components (Low interaction)
  - Suspicious websites are visited with a client honeypot (High interaction)
  - Source code analysis with specialised VM
- Result: 1,287 unique suspicious codes collected, 986 determined to be malicious.

Lit. Review: Fashion Crimes

- First large-scale measurement and analysis of trending-term exploitation.

- Classification of MFA (Made for AdSense) and malware sites.
- Difference between MFA and malware sites tactics.
- Trending-term abuse measurement.
- Economics of exploitation / Revenue analysis.
- Google's intervention reduced revenue by 30%

- Over 60 million search results and tweets collected.
- Result: MFA $100,000 per month, Malware $60,000 per month before search-engine intervention.

Time Issues

- Time cost for data collection in the previous reports:
  - deSEO: 28th May 2010 - 3rd Feb 2011 (~7 months)
  - Automatic Malware Collecting System: 22nd Nov 2010 - 11th Jan 2011 (~2 months)
  - Trending-term Exploitation: 24th Jul 2010 - 24 Apr 2011 (9 months)

- GDP Duration: 1st Oct - 13th Dec 2012 (11 weeks)
  - Start the project from scratch
  - 2 presentations

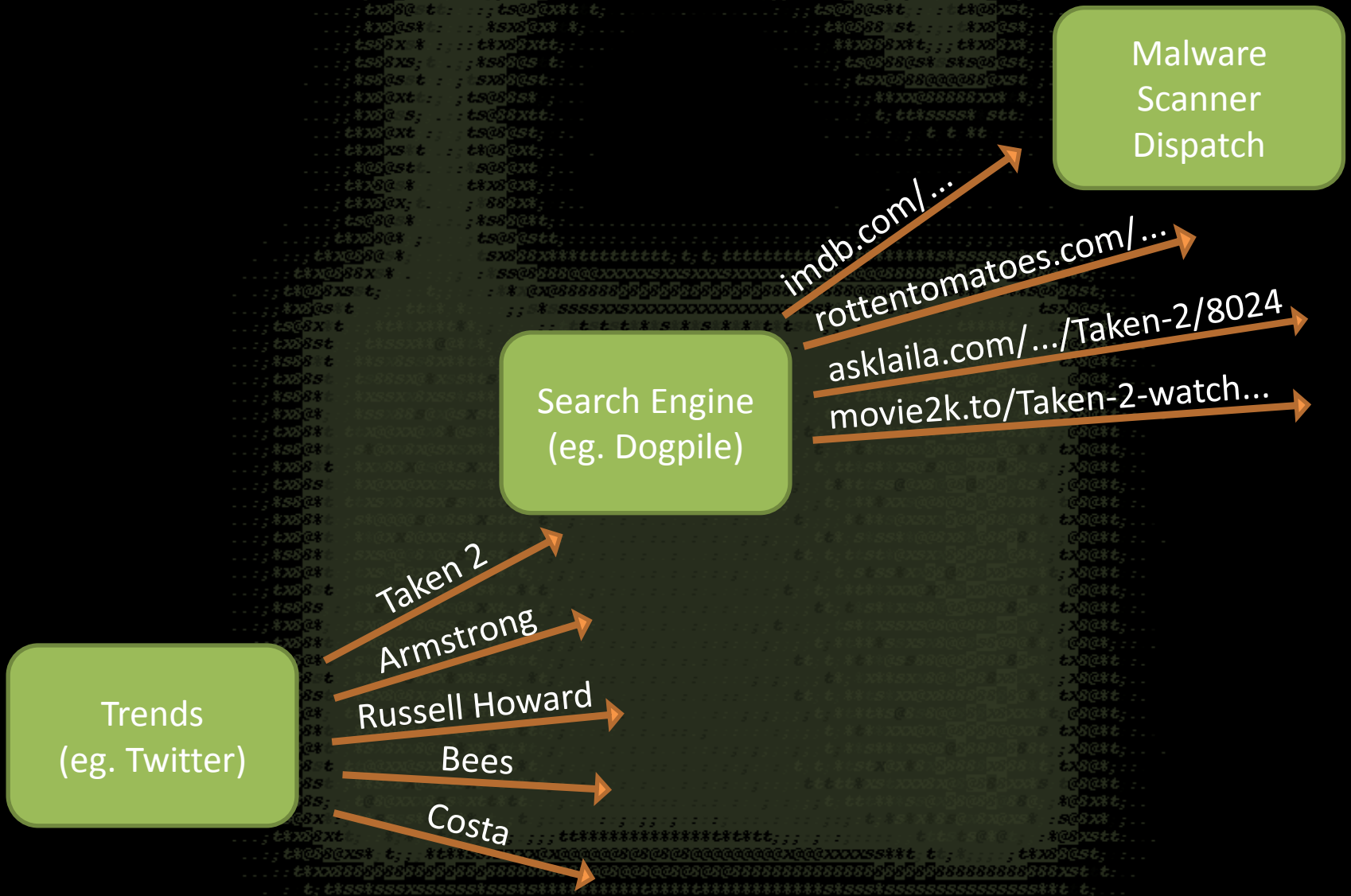
Design: Distributed Architecture

- Celery with RabbitMQ for message passing and Redis to store results
- "Celery is an asynchronous task queue/job queue based on distributed message passing.“
- Work can be distributed over a cluster of many machines on different networks.
- Machines can be added and removed from the cluster without losing tasks

Arch. Design: Task Type Breakdown

- Trend Discovery
  - Find trends on Twitter, or news feeds like The Sun, BBC News and The Register
  - Takes no input, returns a list of keywords
- Search
  - Searches for a keyword in a Search Engine such as Google or Dogpile.
  - Takes a keyword as input, returns a list of URLs
- Malware Scan
  - Queues a page to be scanned for malware.
  - Takes a URL as input, returns a malware report.

Task Dispatch Example



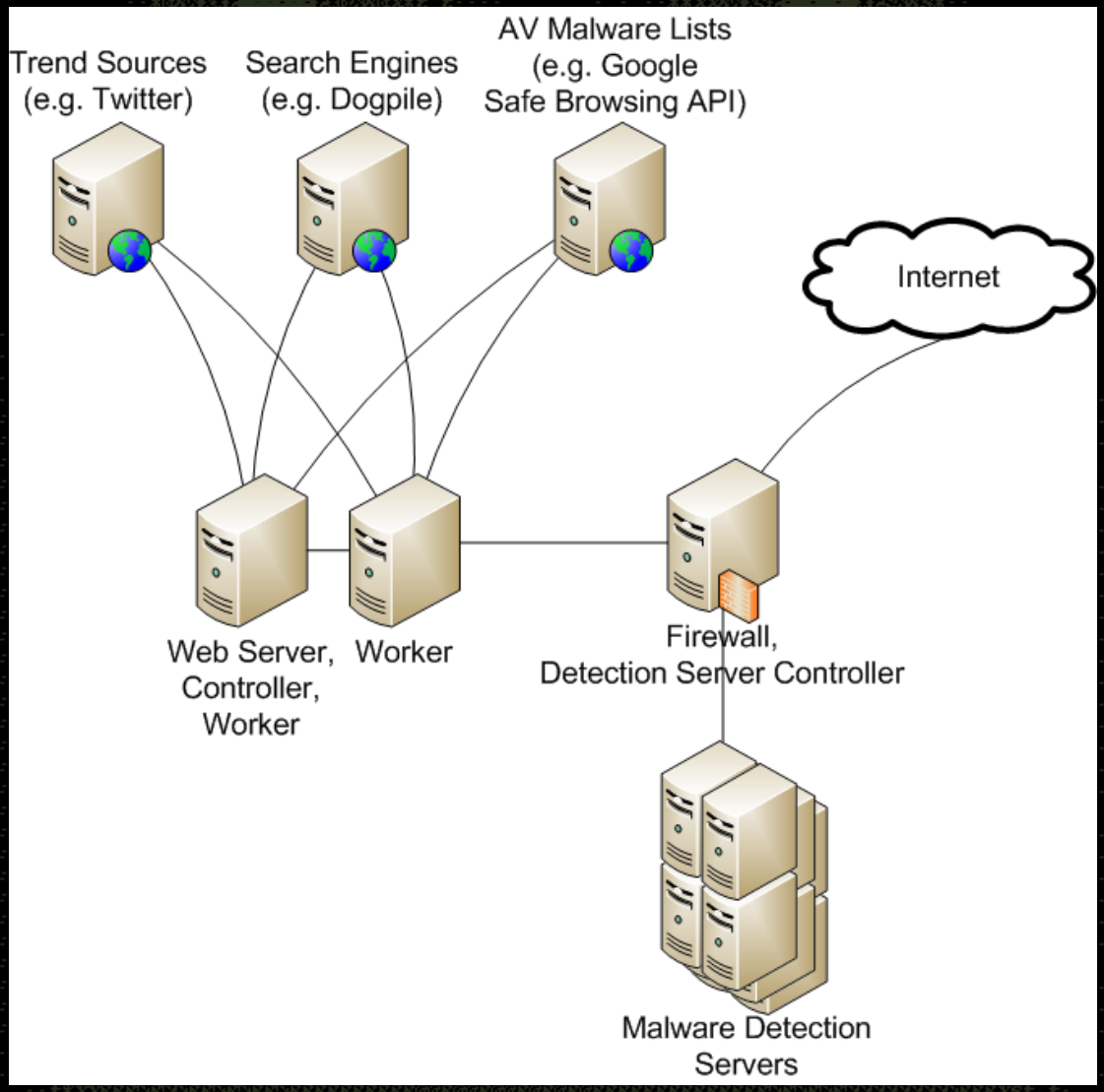
Arch. Design: Periodic Tasks

- Each period run each Trend Discovery, launching Search and Malware Scanning tasks as required.
- Also run a new set of Search and Malware Scanning tasks for each Trend Discovery task that has been run historically.
- So as to investigate malware incidence over time for each trend source
  - How long does it take on average for topics to be targeted by malware?
  - How long before malware is removed?
  - Which source gets the most malware?

Storage & Reporting

- Technologies:
  - Django Web framework
  - Bootstrap Frontend Framework
  - jQuery, AngularJS, D3.js etc...
- Used to:
  - Control and display status of workers
  - Display reports of malware data gathered

Physical Architecture



Malware Detection

- Two main methods for malware detection:
  - Passive malware detection
  - Active malware detection
- Passive method:
  - A malicious attacker injects a malicious code into a user's PC.
- Active method:
  - The malicious code collection system attempt to connect to a particular website and perform malicious action on the website in question. Called client honey pot.

Malware Detection

The client honey pot or active method divided into two groups:

- Low interaction client honey pot
  - Determined what is a malicious webpage
  - The webpage is not rendered and code is not executed
  - The source code of target webpage is downloaded
  - Compare website source with the malicious action pattern of the system
- High interaction client honey pot
  - Render webpage using a web browser
  - Analyse webpage by monitoring malicious behaviour
  - Monitor files, process creation, and registry modification

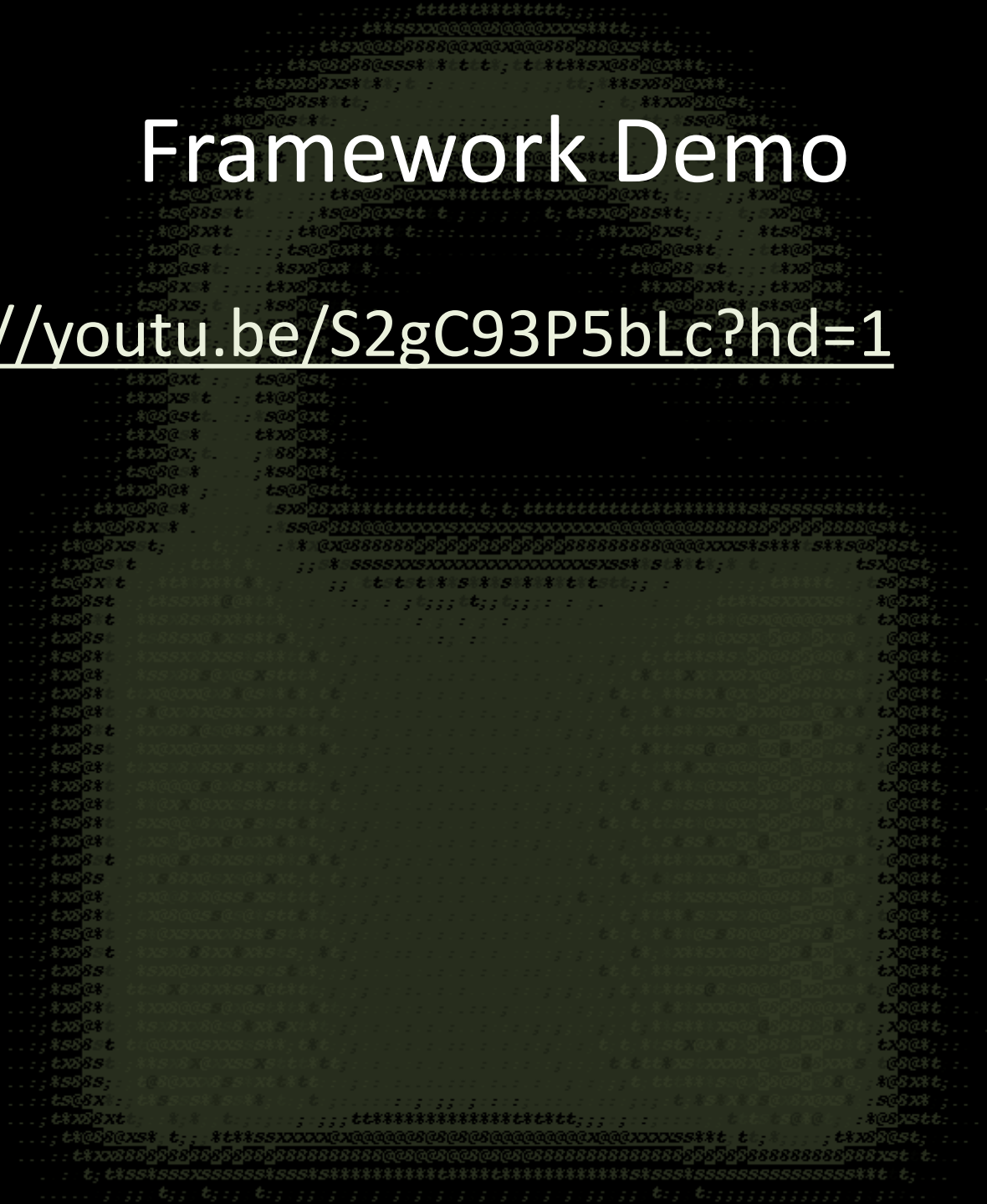
Malware Detection

There two possible solution for malware detection:

- Hybrid client honey pot
- Studying structure of URLs words
  - Studying the structure of URLs contains three steps:
    - Identify suspicious webpage
    - Derive lexical features for each suspicious webpage
    - Perform group analysis to pick out suspicious cluster

Framework Demo

- http://youtu.be/S2gC93P5bLc?hd=1

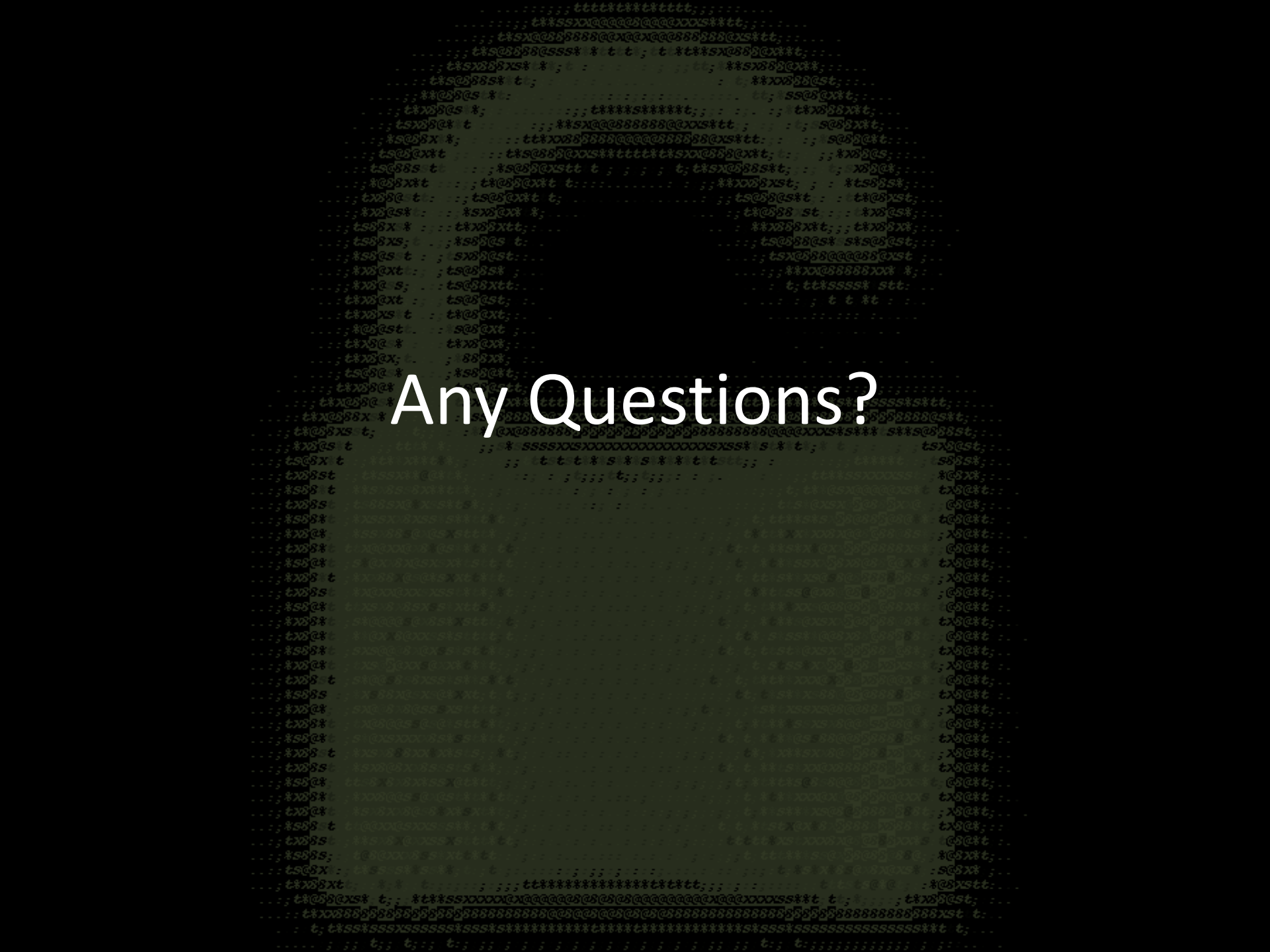


Project Progress

- Architectural Design
- Literature Review
- Development within Framework
  - Twitter, The Sun, and RSS Trend sources
  - Dogpile.co.uk search engine script
  - Placeholder malware anlayser
  - Work on active malware detection
  - Data Model

Remaining Work

- Web Reporting interface
  - Graphing
  - Malware analysis result storage
- Complete active malware detection
- Add passive malware detection
- Compile sample data for report


Any Questions?

References

1. http://www.mcafee.com/us/about/news/2012/q3/20120910-01.aspx?cid=110907
2. John P. John, Fang Yu, Yinglian Xie, Arvind Krishnamurthy, and Martín Abadi. 2011. deSEO: combating search-result poisoning. In *Proceedings of the 20th USENIX conference on Security*(SEC'11). USENIX Association, Berkeley, CA, USA, 20-20.
3. Byung-Ik Kim, B. I. K., Jongil Jeong, J. J., & Hyuncheol Jeong, H. J. (2012). A Study on the Automatic Malware Collecting System Based on the Searching Keyword. *International Journal of Hybrid Information Technology*, *5*(1), 47-60
4. Tyler Moore, Nektarios Leontiadis, and Nicolas Christin. 2011. Fashion crimes: trending-term exploitation on the web. In *Proceedings of the 18th ACM conference on Computer and communications security* (CCS '11). ACM, New York, NY, USA, 455-466.