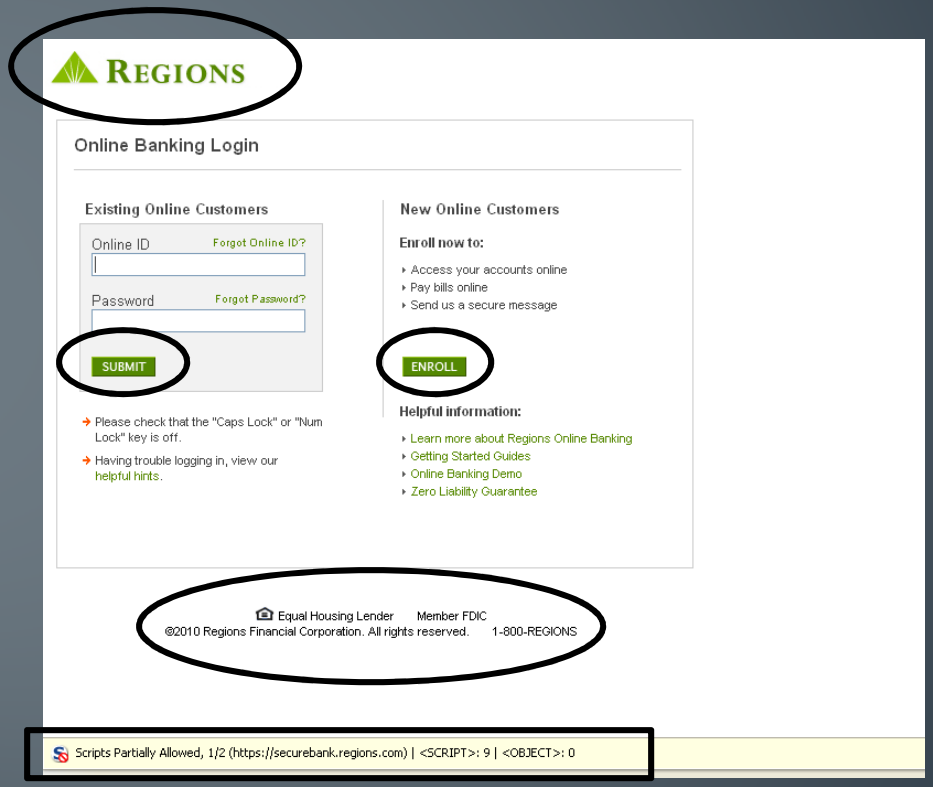
Clustering Potential Phishing Websites Using DeepMD5

Jason Britt

Phishing Sites And Phish Kits

- 41 Files
- 22 Files

Phish File Signatures



Related Work

- Approaches to Identifying Phish
  - Email Based
    - Header Information
    - Email Content
  - URL
  - Content
    - Image Based
    - Source Code Components
  - Mixed
- Clustering Phish By Actors
  - Network Based
    - Domains
    - IP Block
  - Content Based

Overview

- Methods
  - Data Set
  - DeepMD5
  - Slink Clustering Algorithm
  - Associating Phish Kits
- Results
  - Branding
  - Phish Kits
  - Example Cluster Analysis
- Limitations
- Conclusions

Methods: Data Set

- Data from1/1/2011 to 5/25/2011
- Divided into Five Monthly Windows
- Phish Feeds
  - Financial Institutions
  - Security Companies
  - Private Companies
- 265,611 Potential Phishing Websites
  - 38% confirmed phish and branded
  - 12% confirmed non-phish
  - 20% unconfirmed
  - 30% unreachable
- 349 Spoofed Organizations

Methods: Algorithms

- Phase 1 : Main Page Clustering
  - Fingerprint (MD5 hash) Main Page
  - Cluster Websites with the Same Fingerprint

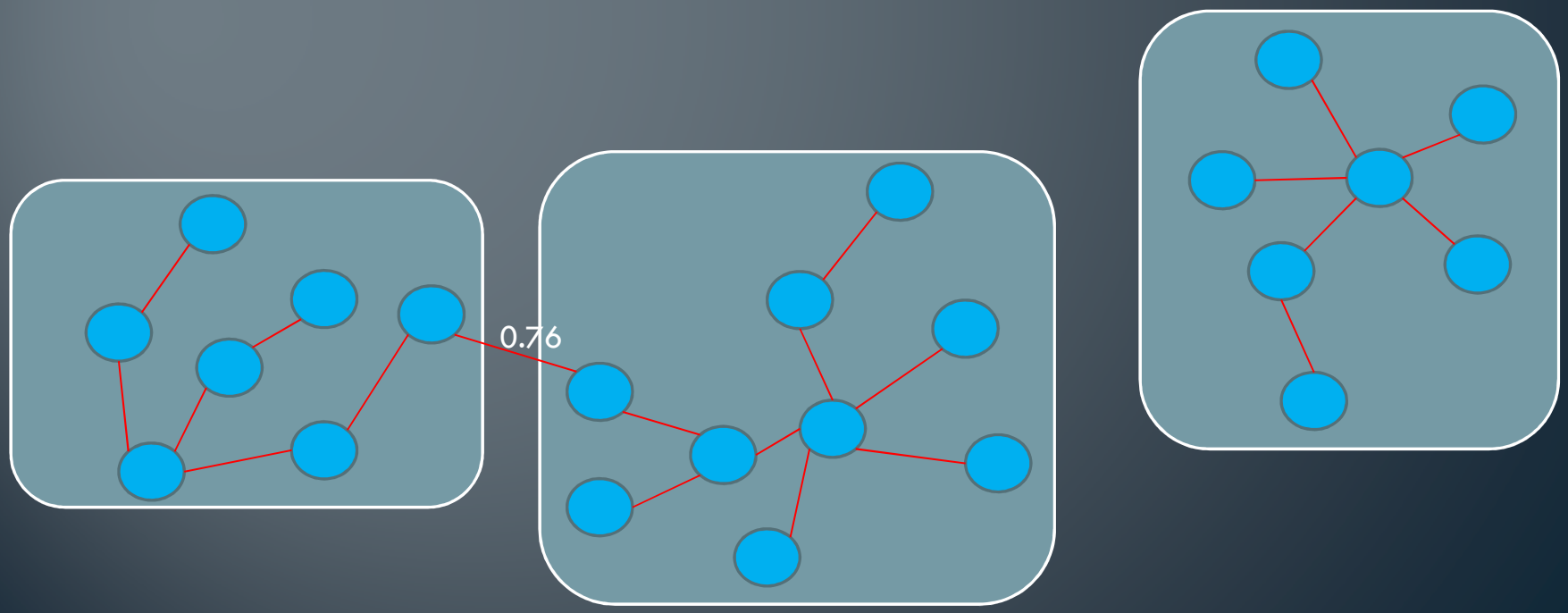
- Phase 2 : SLINK Clustering
  - Fingerprint (MD5 Hash) Support Files
  - DeepMD5 Score > 0.80
  - Cluster Pages using SLINK Algorithm

Methods: DeepMD5

- Set Comparison Coefficient, Kulczynski 2
- Deep MD5 Score $= Average\left(\left(\frac{\text{overlap}}{\text{count1}}\right), \left(\frac{\text{overlap}}{\text{count2}}\right)\right)$
- Examples
  - Website X
    - domain files {a,b,c,d,e}
    - count1 = 5
  - Website Y
    - domain files {a,b,f,g}
    - count2 = 4
  - Overlap = 2
  - Deep MD5 Score = 0.45
    - Average((2/5) , (2/4))

Methods: SLINK Clustering

- Graph Theoretic
  - Phishing Sites are Nodes
  - DeepMD5 Scores are edges
  - Eliminate edges with Low Scores
  - Connected Components are Clusters



Methods: Associating Phish Kits

- Uses Set Comparison Coefficient, Simpson
  - $Simpson = Overlap/Count1$
- Minimum Similarity Threshold 0.80
- Example
  - Phish Site
    - Domain Files {a,b,c}
    - Count1 = 3
  - Phish Kit
    - Kit Files {a,b,c,e,f,g,h}
  - Overlap = 3
  - Simpson Score = 1.0
    - 3/3

Results

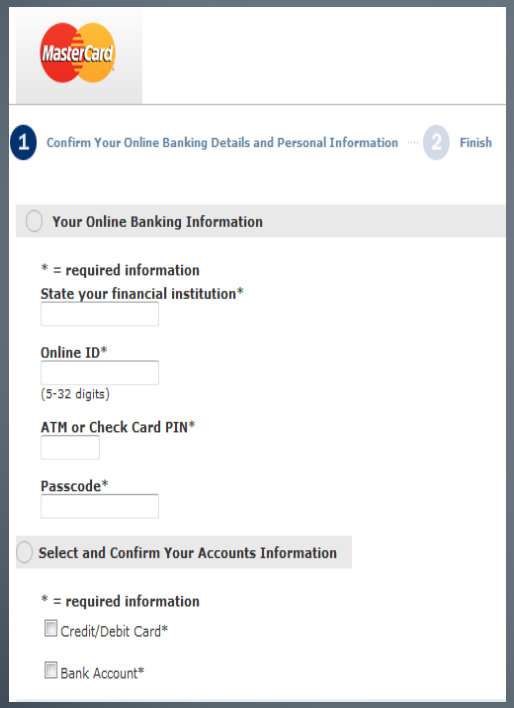
- Methods
  - Data Set
  - DeepMD5
  - Slink Clustering Algorithm
  - Associating Phish Kits
- Results
  - Brands
  - Phishing Kits
  - Example Cluster Analysis
- Limitations
- Conclusions

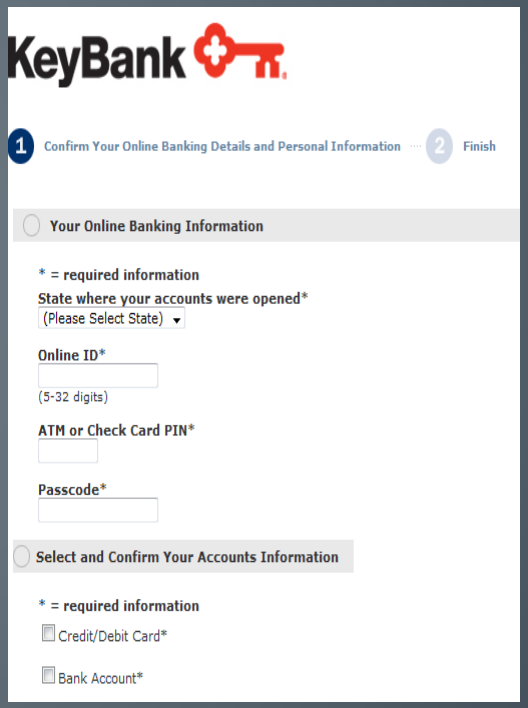
Results: Brands

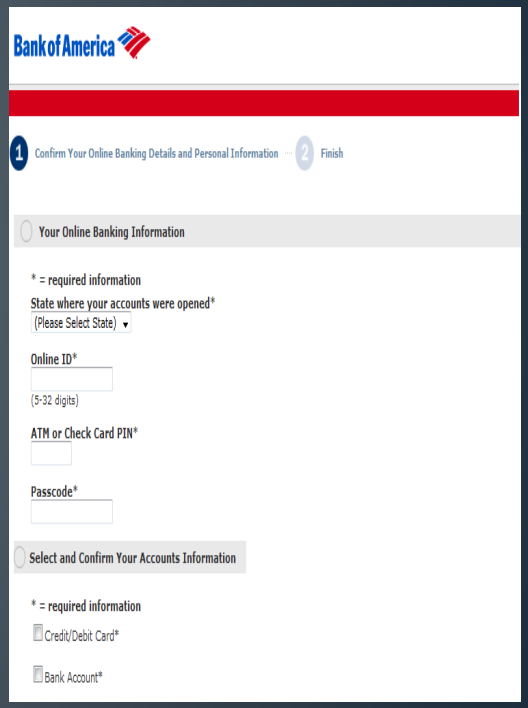
Measure	January	February	March	April	May
Homogeneity	0.9998	0.9994	0.9996	0.9992	0.9989
Completeness	0.5551	0.4123	0.4656	0.4665	0.4812

- 185, 892 Clusters Generated
  - 162,206 Singleton Clusters
  - 22,904 Multi-Member Clusters
- 14,129 Multi-Member Clusters with All Members Branded
  - Size Range : 2 to 1168 Members
  - 14,122 are one brand (pure branded)
  - 7 contain multiple brands (cross-branded)

Results: Cross-Branded Clusters



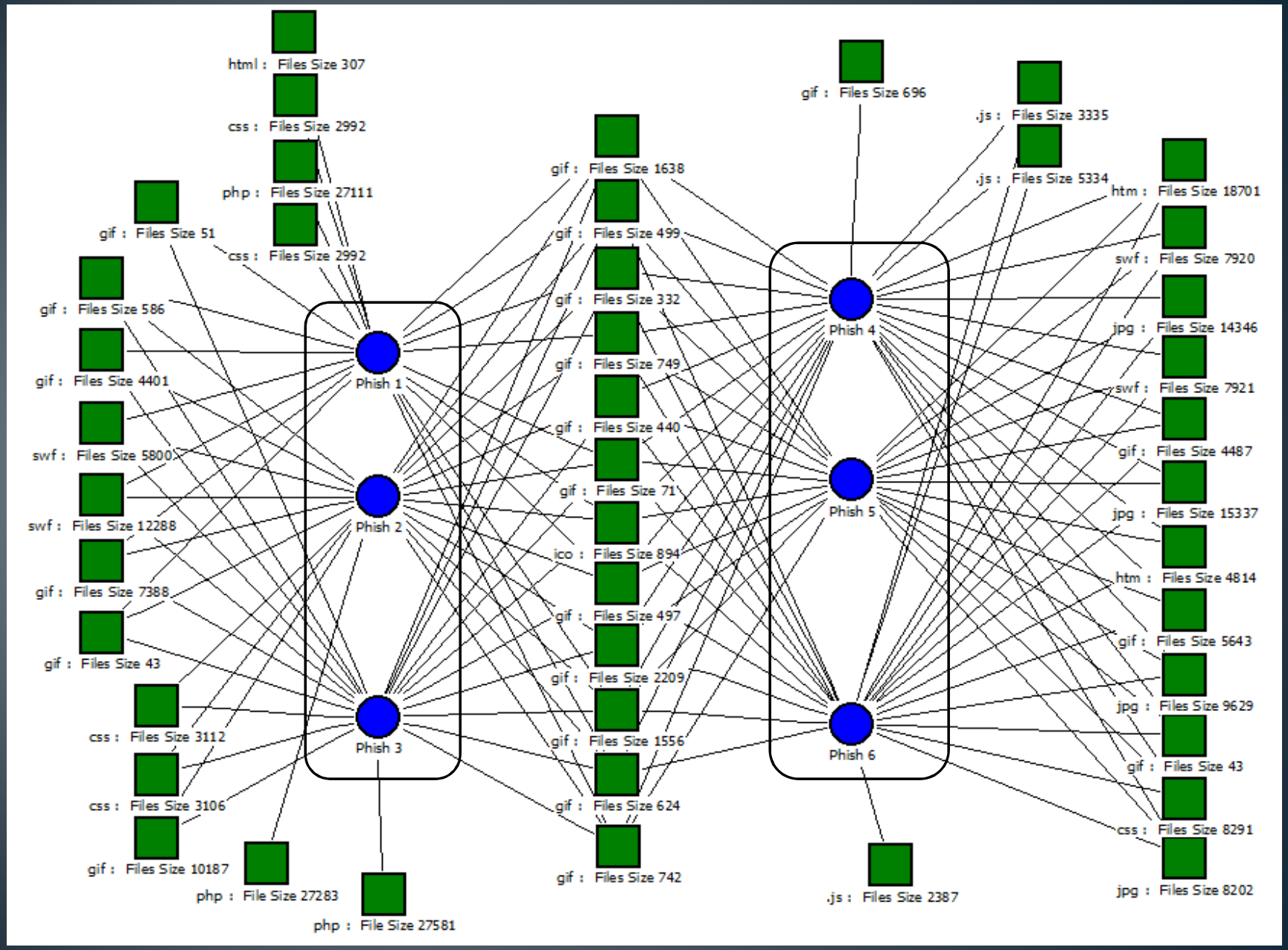


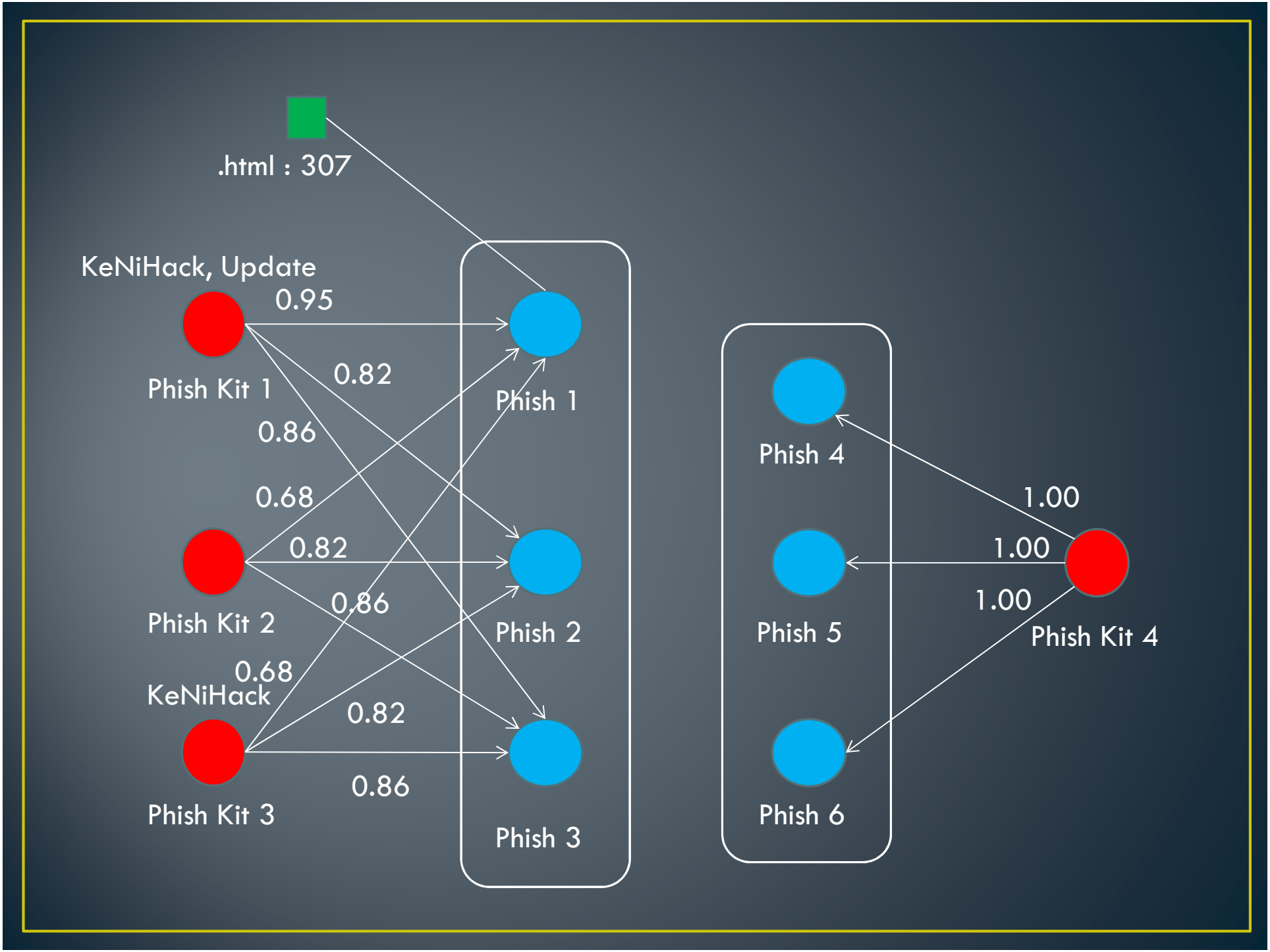


Results: Phishing Kits

Measure	January	February	March	April	May
Cluster Count	2	3	4	1	1
Homogeneity	0.061	0.111	0.138	0.000	0.000
Completeness	1.000	1.000	0.999	1.000	1.000

- 27,801 Phishing Kits to Largest 24 Potential Phish Clusters
- 8,489 Phishing Kits Associated to 6,458 Phish in 11 Phish Clusters
- 6 Phishing Kits Related to > 1 Phish Cluster


html : Files Size 307
css : Files Size 2992
php : Files Size 27111
gif : Files Size 51
css : Files Size 2992
gif : Files Size 586
gif : Files Size 4401
swf : Files Size 5800
swf : Files Size 12288
gif : Files Size 7388
gif : Files Size 43
css : Files Size 3112
css : Files Size 3106
gif : Files Size 10187
php : File Size 27283
php : File Size 27581
Phish 1
Phish 2
Phish 3
gif : Files Size 1638
gif : Files Size 499
gif : Files Size 332
gif : Files Size 749
gif : Files Size 440
gif : Files Size 71
ico : Files Size 894
gif : Files Size 497
gif : Files Size 2209
gif : Files Size 1556
gif : Files Size 624
gif : Files Size 742
gif : Files Size 696
.js : Files Size 3335
.js : Files Size 5334
htm : Files Size 18701
swf : Files Size 7920
jpg : Files Size 14346
swf : Files Size 7921
gif : Files Size 4487
jpg : Files Size 15337
htm : Files Size 4814
gif : Files Size 5643
jpg : Files Size 9629
gif : Files Size 43
css : Files Size 8291
jpg : Files Size 8202
Phish 4
Phish 5
Phish 6
.js : Files Size 2387


.html : 307
KeNiHack, Update
0.95
0.82
0.86
0.68
0.82
0.86
0.68
0.82
0.86
Phish Kit 1
Phish Kit 2
KeNiHack
Phish Kit 3
Phish 1
Phish 2
Phish 3
Phish 4
Phish 5
Phish 6
1.00
1.00
1.00
Phish Kit 4

Html File Redirector

```
<html><head>
<meta http-equiv="refresh" content="0;
URL=Logon.php?LOB=RBGLogon&_pageLabel=page_logonform">
<script language="JavaScript" type="text/javascript">
<!--
function redirect() {
setTimeout("window.location.replace('Logon.php?LOB=RBGLogon&_pag
eLabel=page_logonform')", 0); }
-->
</script>
</head>
```

Conclusions

- Clusters Evaluated By Brand
  - Highly Homogenous
  - Multiple Large Clusters With The Same Brand
- Clusters Evaluated By Associated Phish Kits
  - Associating Phish Kits To Phish Cluster Members
    - Limited Support
    - Many Phish Kits Relate To One Phish Cluster
  - Example Analysis
    - Limited Support
    - Phish Kits Relating To A Phish Cluster Are Related
- Phish are Clustered by Phish Kit Family

Limitations

- Innocuous Files
  - Examples
    - Single pixel Image Files
    - Common Web Statistics Files
  - Diluting Structural Comparison Effect
- Small File Count Websites
  - Deep MD5 Comparison Problems
  - Retrieve Only Local Domain Files
- Clusters Not necessarily based upon Similar “Look And Feel”
  - Based upon Structure
  - Similar Structure = Similar “Look And Feel”
  - Similar “Look And Feel” != Similar Structure

Future Work

- Syntactical Fingerprinting
  - Uses Only Main Html Page
  - Breaks Main Page into Multiple Components
  - May Alleviate Small File Count Problem
  - May Alleviate Innocuous File Problem
- Phish Kit Clustering
  - Cluster Phish Kits Using DeepMD5 SLINK
  - Relate Phish Kit Clusters to Phish Clusters
  - Reduce Run Time of Relating Phish Kits to Phish Clusters

Special Thanks

- Dr. Alan Sprague
- Gary Warner
- Brad Wardman
- UAB Phishing Operations Team