CE 817 - Advanced Network Security
Phishing II

Lecture 16

Mehdi Kharrazi
Department of Computer Engineering
Sharif University of Technology

Acknowledgments: Some of the slides are fully or partially obtained from other sources. Reference is noted on the bottom of each slide, when the content is fully obtained from another source. Otherwise a full list of references is provided on the last slide.

CANTINA: A Content-Based Approach to Detecting Phishing Web Sites

Yue Zhang, University of Pittsburgh, Jason I. Hong, Lorrie F. Cranor
Carnegie Mellon University, www2007.

Strategies to Counter Phishing

- Make it invisible
  - Taking down phishing web pages
  - Filtering out phishing email
  - Detecting phishing web pages (SpoofGuard, etc)
- Provide better user interfaces
  - Extended certificate verification
  - Anti-phishing toolbars (SpoofGuard, eBay, Netcraft, etc)
- Train the users
  - Games (Sheng et al, SOUPS 2007)

Two Ways of Detecting Phishing Pages

- Human-verified Blacklists
  - No false positives, easy to implement, robust to new attacks
  - But tedious, slow to update, and not comprehensive
  - Only one toolbar found more than 60% phishing sites (Egelman et al, NDSS 2007)

- Heuristics
  - Fast to find new phishing sites (zero-day)
  - But false positives, may be fragile to new attacks
  - Not much work in this area
  - Our work contributes to the understanding of heuristics

Our Solution: CANTINA

- CANTINA uses a simple content-based approach
  - Examines content of a web page and creates a “fingerprint”
  - Sends that fingerprint as a query to a search engine
  - Sees if the web page in question is in the top search results
    - If so, then we label it legitimate
    - Otherwise, we label it phishing

- Nice properties:
  - Fast
  - Scales well
  - No maintenance by us (done by search engines)
  - Highly accurate

How Robust Hyperlinks Work

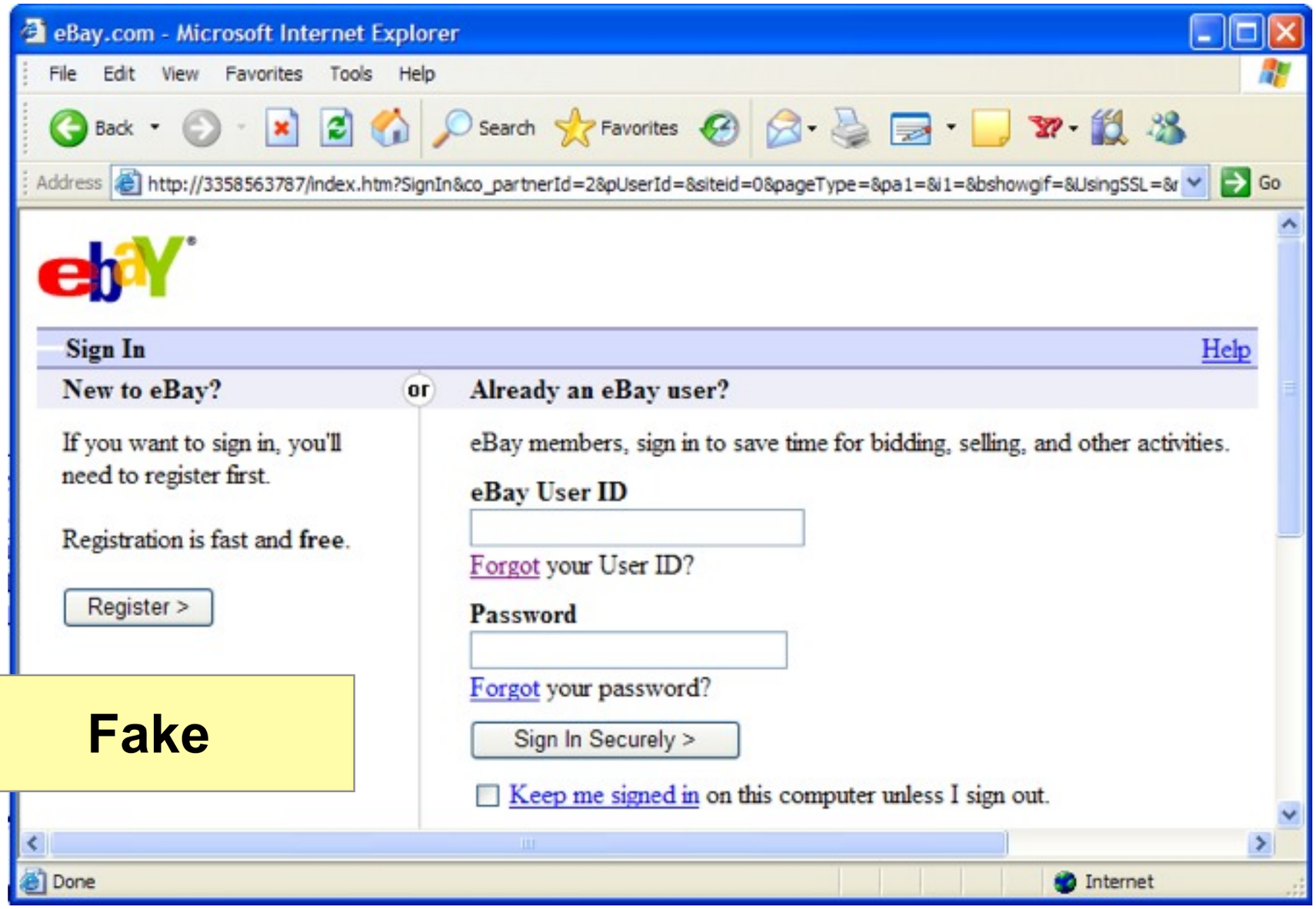
- Developed by Phelps and Wilensky to solve
  “404 not found” problem (D-Lib Magazine 2000)

- Add lexical signature to URLs
  - If link doesn’t work, then feed signature to search engine
  - Ex. http://abc.com/page.html?sig=“word1+word2+...+word5”

- How to generate useful signatures?
  - Term Frequency / Inverse Document Frequency (TF-IDF)
  - Their informal evaluation found using top five words
    as scored by TF-IDF was surprisingly effective

Adapting TF-IDF for Anti-Phishing

- Can same basic approach be used for anti-phishing?
  - Scammers often directly copy legitimate web pages or include keywords like name of legitimate organization

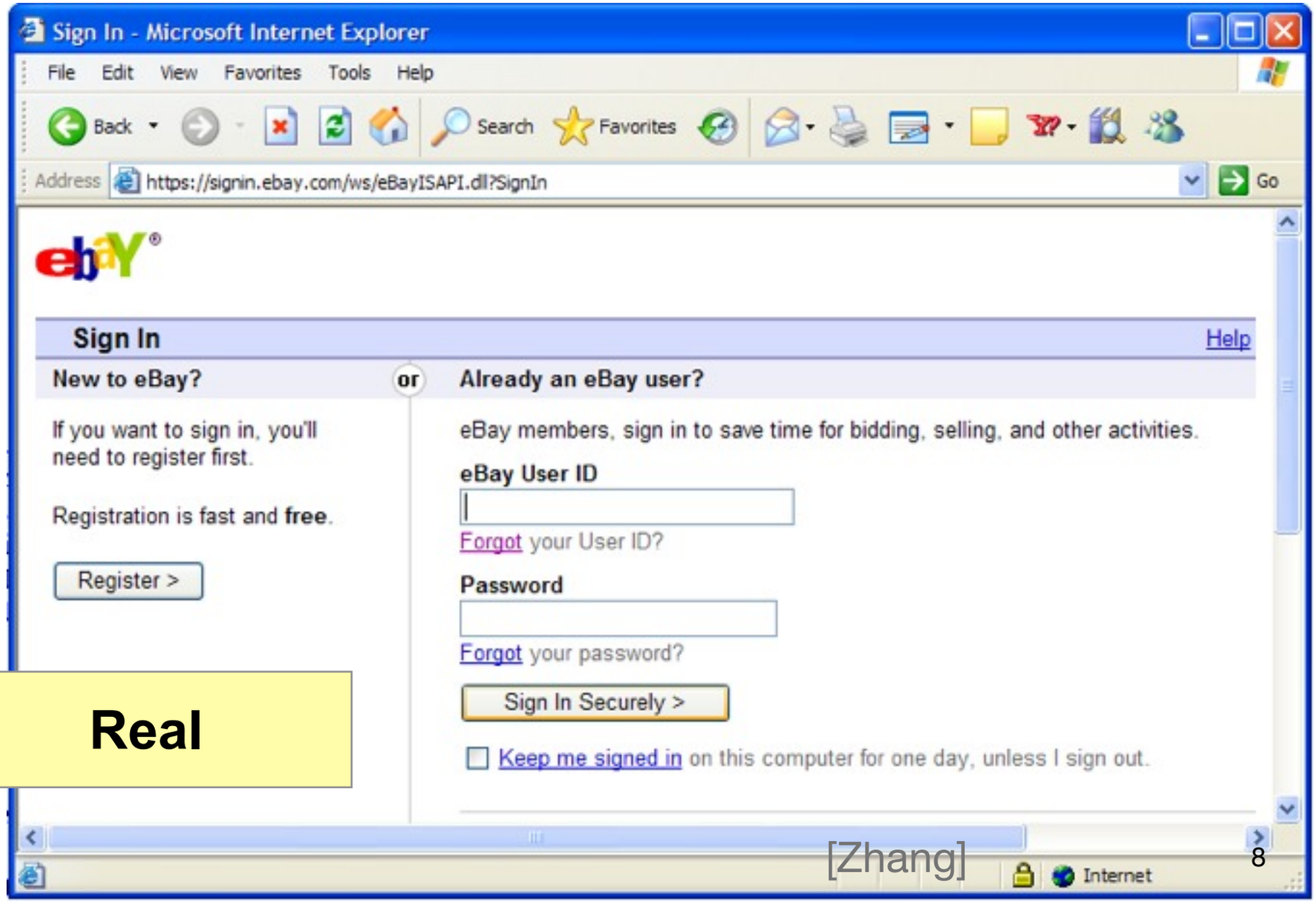


Fake

[Zhang]

Adapting TF-IDF for Anti-Phishing

- Can same basic approach be used for anti-phishing?
  - Scammers often directly copy legitimate web pages or include keywords like name of legitimate organization



Real

[Zhang]

Adapting TF-IDF for Anti-Phishing

- Can same basic approach be used for anti-phishing?
  - Scammers often directly copy legitimate web pages or include keywords like name of legitimate organization
  - With Google, phishing site should have low page rank
    - APWG states that phishing sites alive 4.5 days
    - Few sites link to phishing sites
    - Hence, phishing sites unlikely to be in top search results

- Hypothesis:
  - CANTINA will be able to discriminate between legitimate and phishing sites quite well

How CANTINA Works (Iteration #1)

- Given a web page, calculate TF-IDF score for each word in that page
- Take five words with highest TF-IDF weights
- Feed these five words into a search engine (Google)
- If domain name of current web page is in top N search results, we consider it legitimate
  - N=30 worked well
  - No improvement by increasing N

Fake

eBay.com - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://3358563787/index.htm?SignIn&co_partnerId=2&pUserId=&siteid=0&pageType=&pa1=&i1=&bshowgif=&UsingSSL=&r Go

eBay, user, sign, help, forgot

Sign In Help

New to eBay? or **Already an eBay user?**

If you want to sign in, you'll need to register first.

Registration is fast and **free**.

Register >

eBay members, sign in to save time for bidding, selling, and other activities.

eBay User ID

Forgot your User ID?

Password

Forgot your password?

Sign In Securely >

Keep me signed in on this computer unless I sign out.

Done Internet

[Zhang]

Real

Sign In - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address https://signin.ebay.com/ws/eBayISAPI.dll?SignIn Go

eBay, user, sign, help, forgot

ebay®

Sign In Help

New to eBay? or **Already an eBay user?**

If you want to sign in, you'll need to register first.

Registration is fast and **free**.

Register >

eBay members, sign in to save time for bidding, selling, and other activities.

eBay User ID

Forgot your User ID?

Password

Forgot your password?

Sign In Securely >

Keep me signed in on this computer for one day, unless I sign out.

Internet

[Zhang]

Google

Web Images Video News Maps more »

eBay, user, sign, help, forgot | Search | Advanced Search | Preferences

Web Results **1** - **10** of about **4,080,000** for **eBay**, **user**, **sign**, **help**, **forgot**. (**0.24** seconds)

Forgot User ID
Hello! **Sign** in/out. **Forgot User** ID **...** Use of this Web site constitutes acceptance of the **eBay User** Agreement and Privacy Policy. **...**
cgi4.**ebay**.com/ws/**eBay**ISAPI.dll?**User**IdRecognizerShow - 16k - Cached - Similar pages

About Signing in to Your Account
If you ever **forget** your **User** ID or password, you can have **eBay** send it to your **...** If you're a Passport **user**, **sign** in with your **eBay User** ID and password. **...**
pages.**ebay**.com/**help**/newto**ebay**/**sign**in.html - 33k - Cached - Similar pages

Problems Signing-in?
eBay uses cookies to **help** you **sign** in. Make sure that your browser is set to **...** I cannot **sign** in because I **forgot** my **User** ID. Ask **eBay** to send your **User** ID **...**
pages.**ebay**.com/**help**/newto**ebay**/**sign**-in-trouble.html - 35k - Cached - Similar pages
[More results from pages.ebay.com]

Welcome - PayPal
Member Log-In, **Forgot** your email address? **Forgot** your password? **...** Free **eBay** tools make selling easier. PayPal works hard to **help** protect sellers. **...**
www.paypal.com/ - 21k - Cached - Similar pages

Sign In
Sign In, **Help ...** If you want to **sign** in, you'll need to register first. **...** Use of this Web site constitutes acceptance of the **eBay User** Agreement and **...**
signin.**ebay**.com/ws/**eBay**ISAPI.dll?**Sign**In - 16k - Cached - Similar pages

Sponsored Links

eBay ® - Official Site
Looking for **eBay**?
Find exactly what you want today.
www.**ebay**.com

eBay Selling **Help**
Full-service **eBay** selling solution.
Easy, convenient & great results.
SellingRequest.com
Philadelphia, PA

[Zhang]

Web Images Video News Maps more »

eBay, user, sign, help, forgot

Sear

Web

econds)

Forgot User ID
Hello! **Sign** in/out. **Forgot User** ID **...** Use of this Web site constitutes acceptance of the **eBay User** Agreement and Privacy Policy. **...**
cgi4.**ebay**.com/ws/**eBay**ISAPI.dll?**User**IdRecognizerShow - 16k - Cached - Similar pages

About Signing in to Your Account
If you ever **forget** your **User** ID or password, you can have **eBay** send it to your **...** If you're a Passport **user**, **sign** in with your **eBay User** ID and password. **...**
pages.**ebay**.com/**help**/newto**ebay**/**sign**in.html - 33k - Cached - Similar pages

Problems Signing-in?
eBay uses cookies to **help** you **sign** in. Make sure that your browser is set to **...** I cannot **sign** in because I **forgot** my **User** ID. Ask **eBay** to send your **User** ID **...**
pages.**ebay**.com/**help**/newto**ebay**/**sign**-in-trouble.html - 35k - Cached - Similar pages
[More results from pages.ebay.com]

Welcome - PayPal
Member Log-In, **Forgot** your email address? **Forgot** your password? **...** Free **eBay** tools make selling easier. PayPal works hard to **help** protect sellers. **...**
www.paypal.com/ - 21k - Cached - Similar pages

Sign In
Sign In, **Help ...** If you want to **sign** in, you'll need to register first. **...** Use of this Web site constitutes acceptance of the **eBay User** Agreement and **...**
signin.**ebay**.com/ws/**eBay**ISAPI.dll?**Sign**In - 16k - Cached - Similar pages

Sponsored Links

eBay ® - Official Site
Looking for **eBay**?
Find exactly what you want today.
www.**ebay**.com

eBay Selling **Help**
Full-service **eBay** selling solution.
Easy, convenient & great results.
SellingRequest.com
Philadelphia, PA

[Zhang]

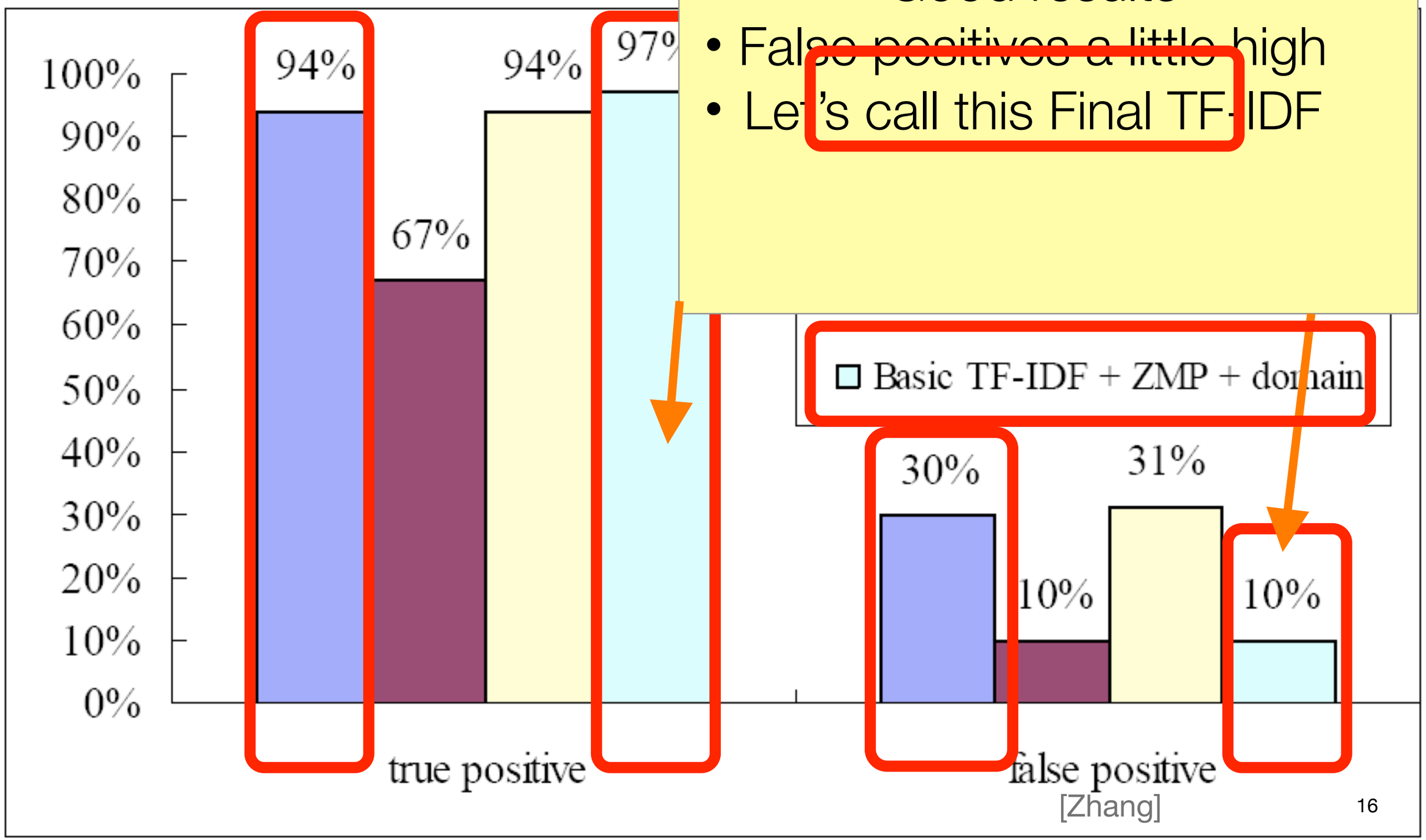
Evaluating CANTINA (Iteration #1)

- 100 phishing URLs from PhishTank.com
  - We used unverified URLs, manually verified them ourselves
- 100 legitimate URLs from another study on phishing
  - From 3Sharp, popular web sites, banks, etc

- Four conditions
  - Basic TF-IDF
  - Basic TF-IDF + domain name (ebay.com –> “ebay”)
  - Basic TF-IDF + ZMP (zero results means phishing)
  - Basic TF-IDF + domain name + ZMP

Evaluating CANTINA (Iteration #1)

- Good results
- False positives a little high
- Let's call this Final TF-IDF



[Zhang]

How CANTINA Works (Iteration #2)

- Wanted to reduce false positives
- Added several heuristics from SpoofGuard and PILFER
  - Age of domain
  - Known images (logos)
  - Page is at suspicious URL (has @ or -)
  - Page contains suspicious links
  - IP Address in URL
  - Dots in URL (>= 5 dots)
  - Page contains text entry fields
  - TF-IDF

How CANTINA Works (Iteration #2)

- Used simple forward linear model to weight these
  - The more effective a heuristic, the larger the weight
  - Used 100 phishing URLs, 100 legitimate to find weights

Heuristic	True Positive	False Positive	Effect	Weight
Age of Domain	87%	30%	57.0	0.18
Known Images	37%	0%	37.0	0.12
Suspicious URL	6%	3%	3.0	0.01
Suspicious Links	8%	25%	0.0	0.00
IP Address	22%	0%	22.0	0.07
Dots in URL	45%	3%	42.0	0.13
Forms	94%	27%	67.0	0.21
TF-IDF-Final	99%	10%	89.0	0.28

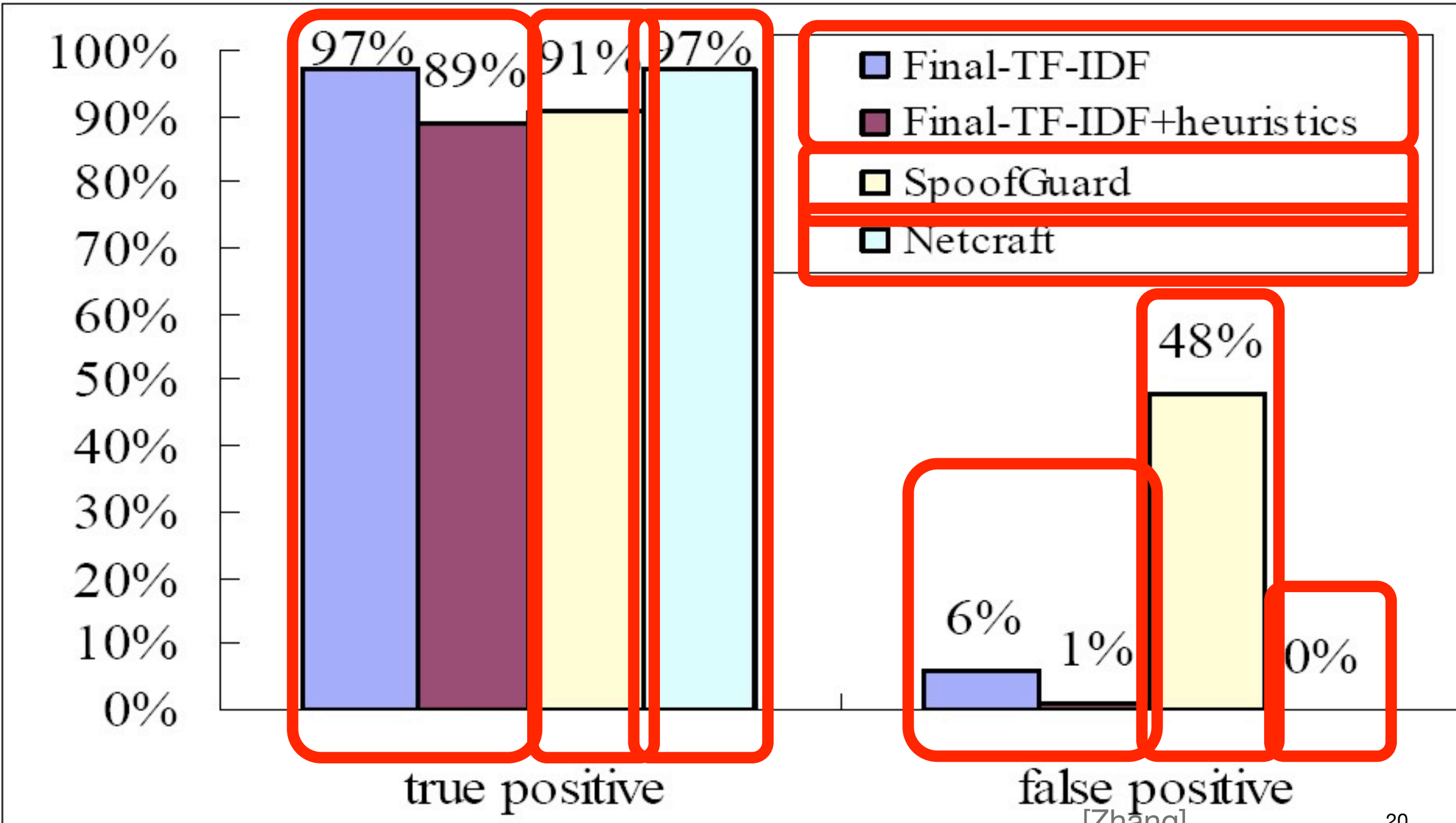
[Zhang]

Evaluating CANTINA (Iteration #2)

- Compared CANTINA to SpoofGuard and NetCraft
  - SpoofGuard uses all heuristics
  - NetCraft 1.7.0 uses heuristics and extensive blacklist

- 100 phishing URLs from PhishTank.com
- 100 legitimate URLs
  - 35 sites often attacked (citibank, paypal)
  - 35 top pages from Alexa (most popular sites)
  - 30 random web pages from random.yahoo.com

Evaluating CANTINA (Iteration #2)



[Zhang]

Discussion of Evaluation

- Good results again for CANTINA (iteration #2)
  - 97% with 6% false positive, 89% with 1% false positive
- CANTINA close to Netcraft (human-verified)
- Shifts problem of identifying phishing sites to a search engine problem

Discussion of CANTINA Overall

- Limitations
  - Does not work well for non-English web sites (TF-IDF)
  - System performance (querying Google each time)
    - Early results from our latest work => low latency crucial
    - CANTINA may be better for backend work than browser
- Attacks by criminals
  - Using images instead of words
    - But has to look legitimate
  - Invisible text
  - But phishing page still has to be in top search results
    - Circumventing TF-IDF and PageRank (hard in practice?)

An Inquiry into the Nature and Causes of the Wealth of Internet Miscreants

Jason Franklin, Vern Paxon, Adrian Perrig, and Stefan Savage

Commoditization of eCrime

Shift from Hacking For Fun to For Profit

- Observation 1: Internet-based crime shifting from reputation economy to cash economy
  - Today, large fraction of Internet-based crime is profit driven
  - Can be modeled roughly as rational behavior
- Observation 2: eCrime has expanded and evolved to exceed capacity of closed group
  - There now exists diverse on-line market economy that trades in illicit goods and services in support of criminal activity

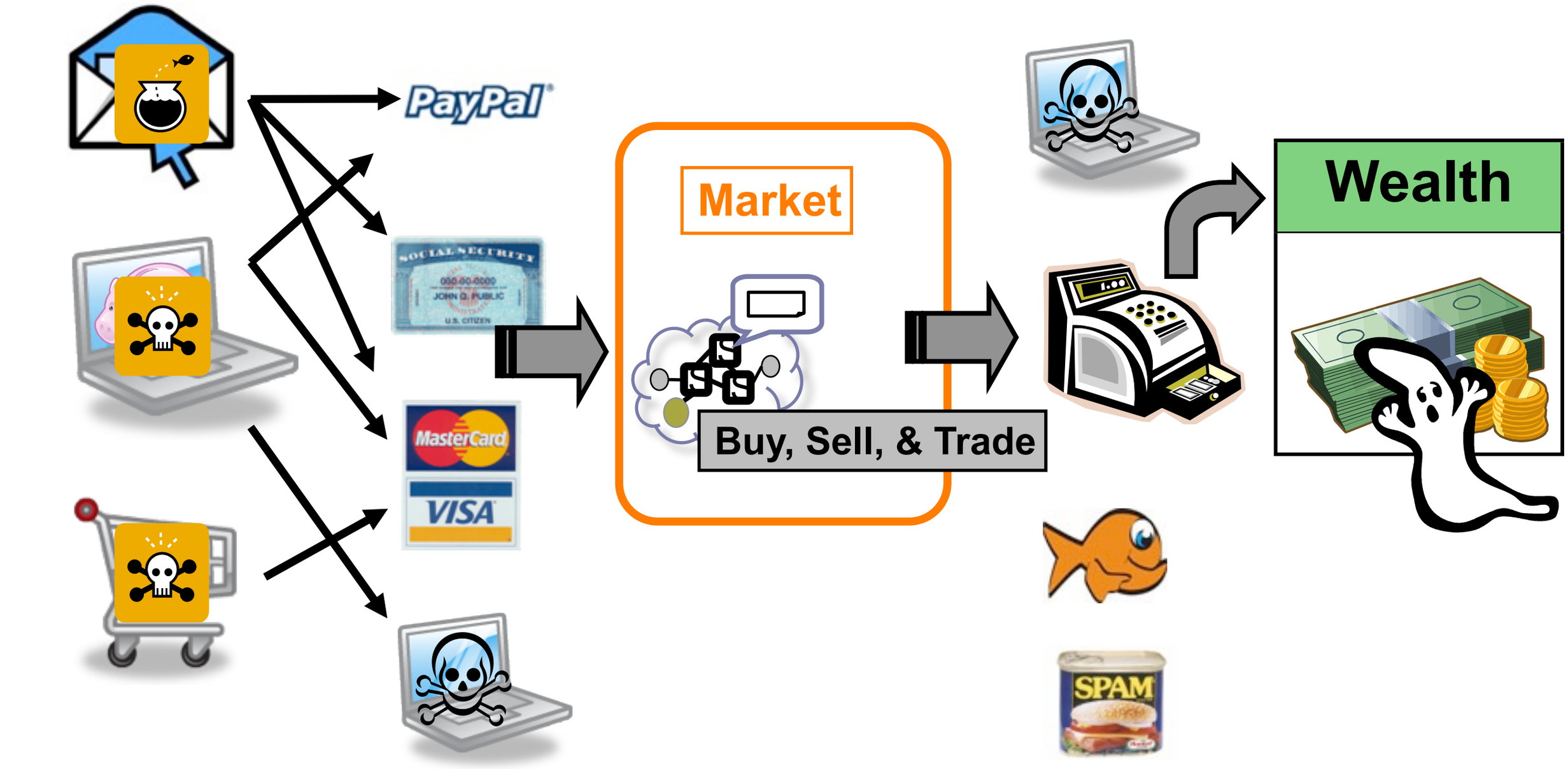
- Markets are public, bustling with activity, easy to access
  - Lower barrier to entry for eCrime, increase profitability, and contribute to overall level of Internet-based criminal activity

Contributions

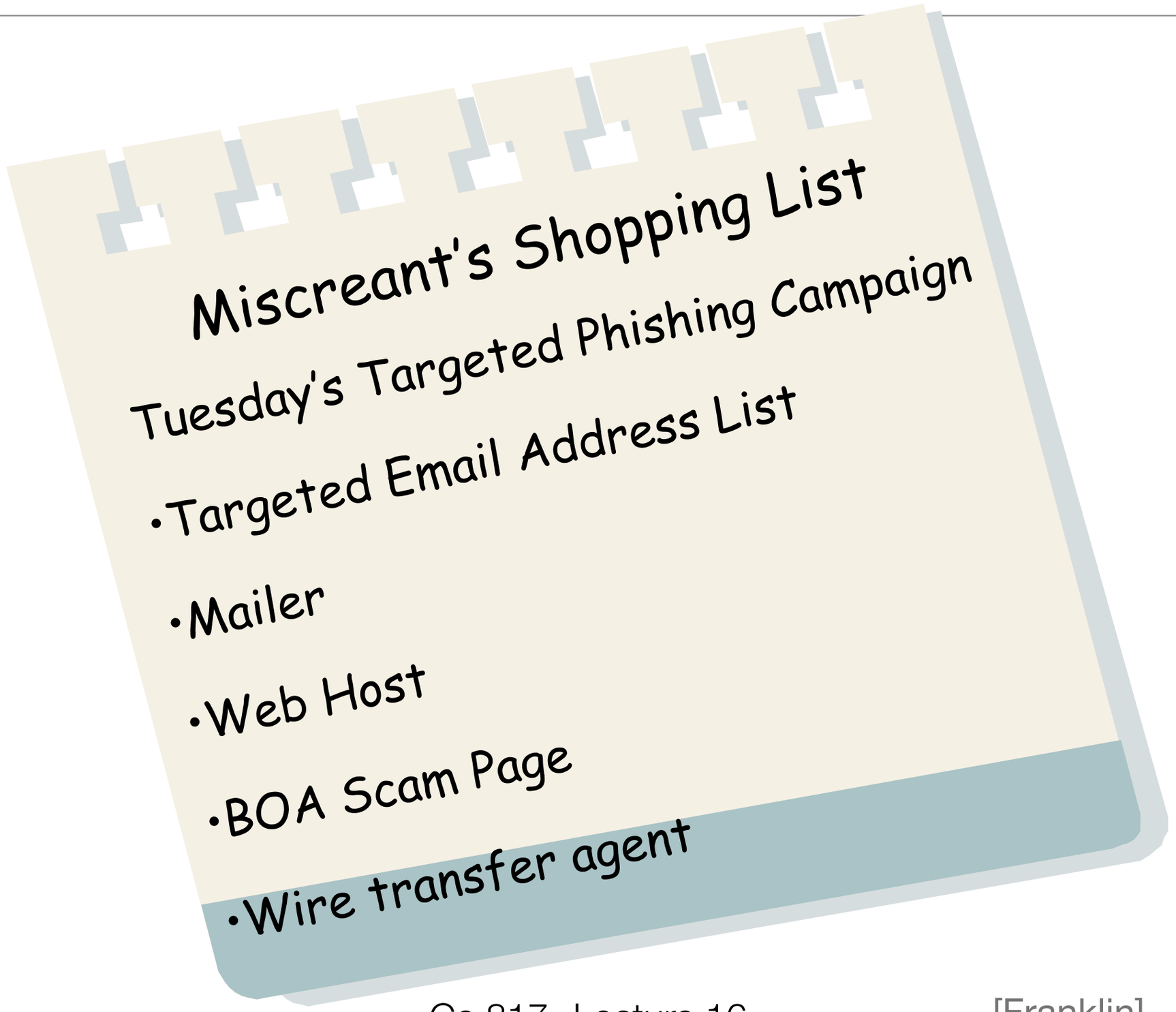
- First systematic exploration into measuring and analyzing eCrime market
- Characterize participants and explore goods and services offered
- Discuss beneficial uses of market data
- Discuss market disruption

eCrime Market Operation

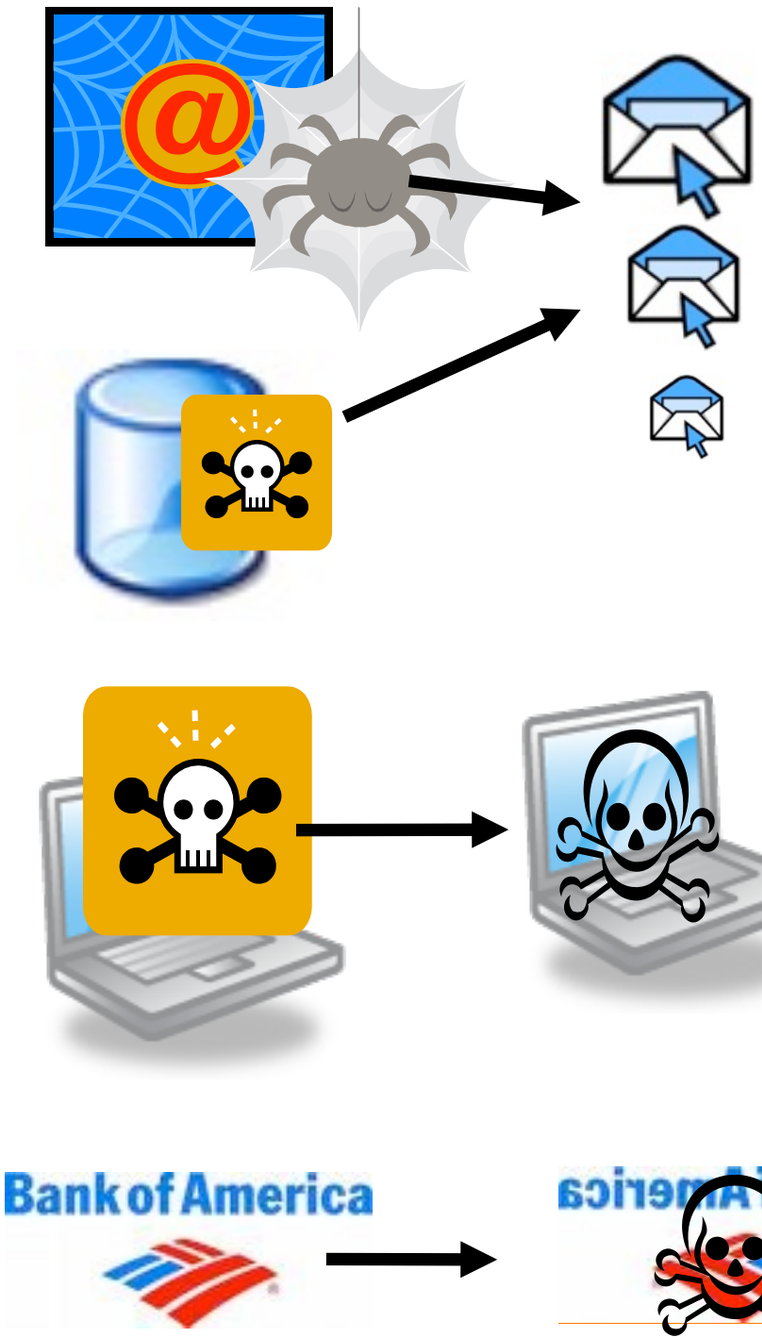
Raw Materials | Goods | (Re)Application | Goal



Market Lowers Barrier to Entry for eCrime

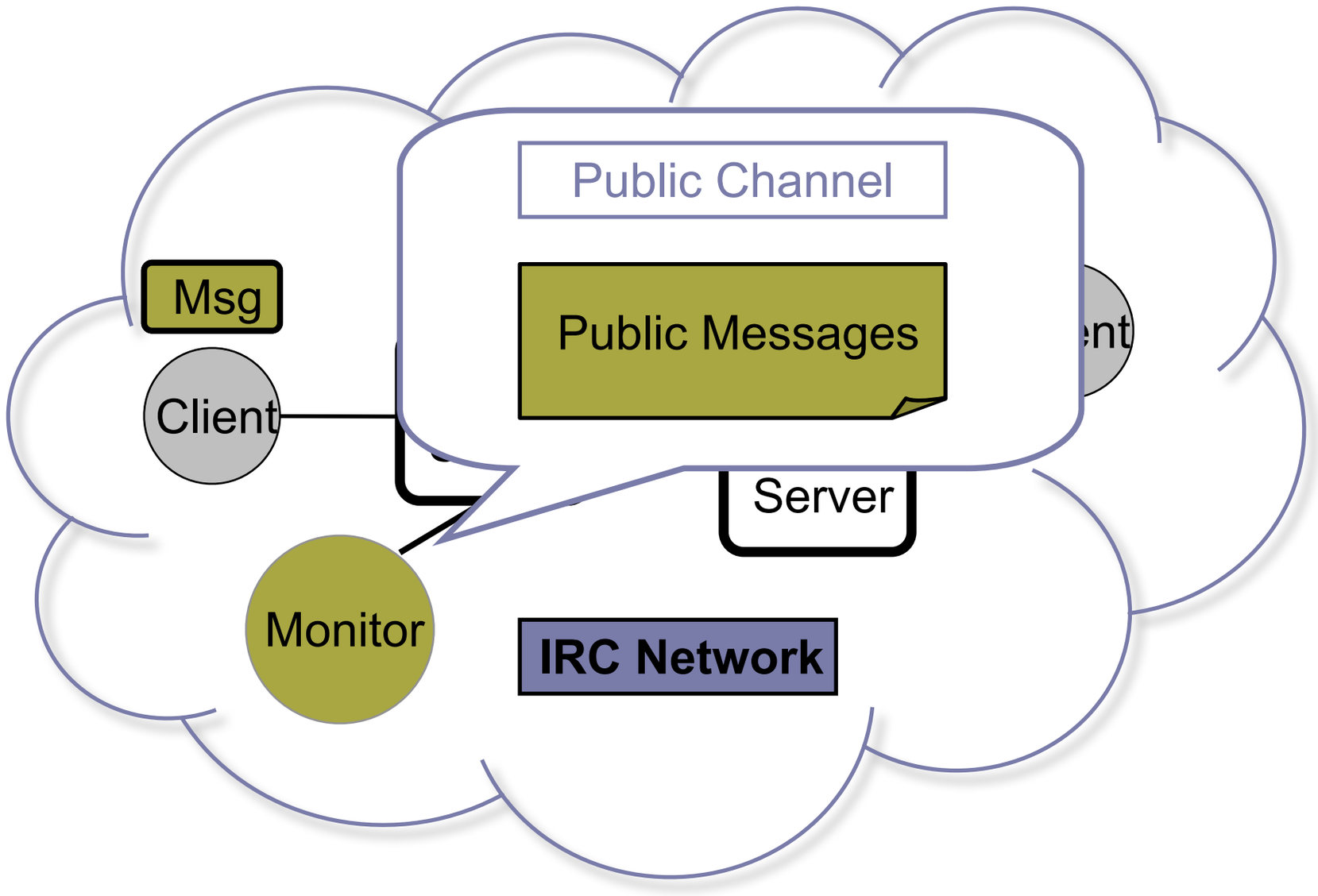


Buying a Targeted Phishing Campaign



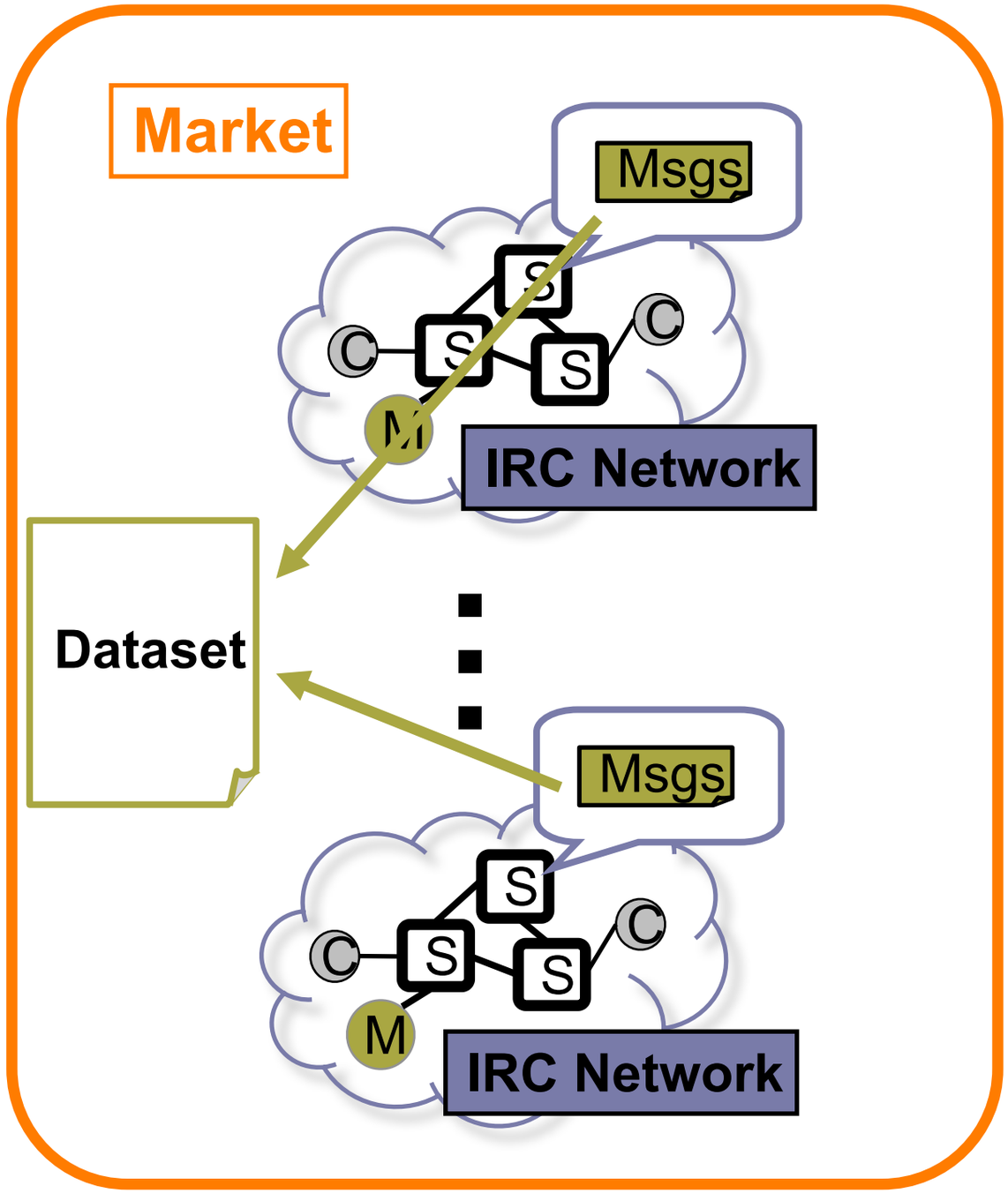
[Franklin]

Market Data Collection



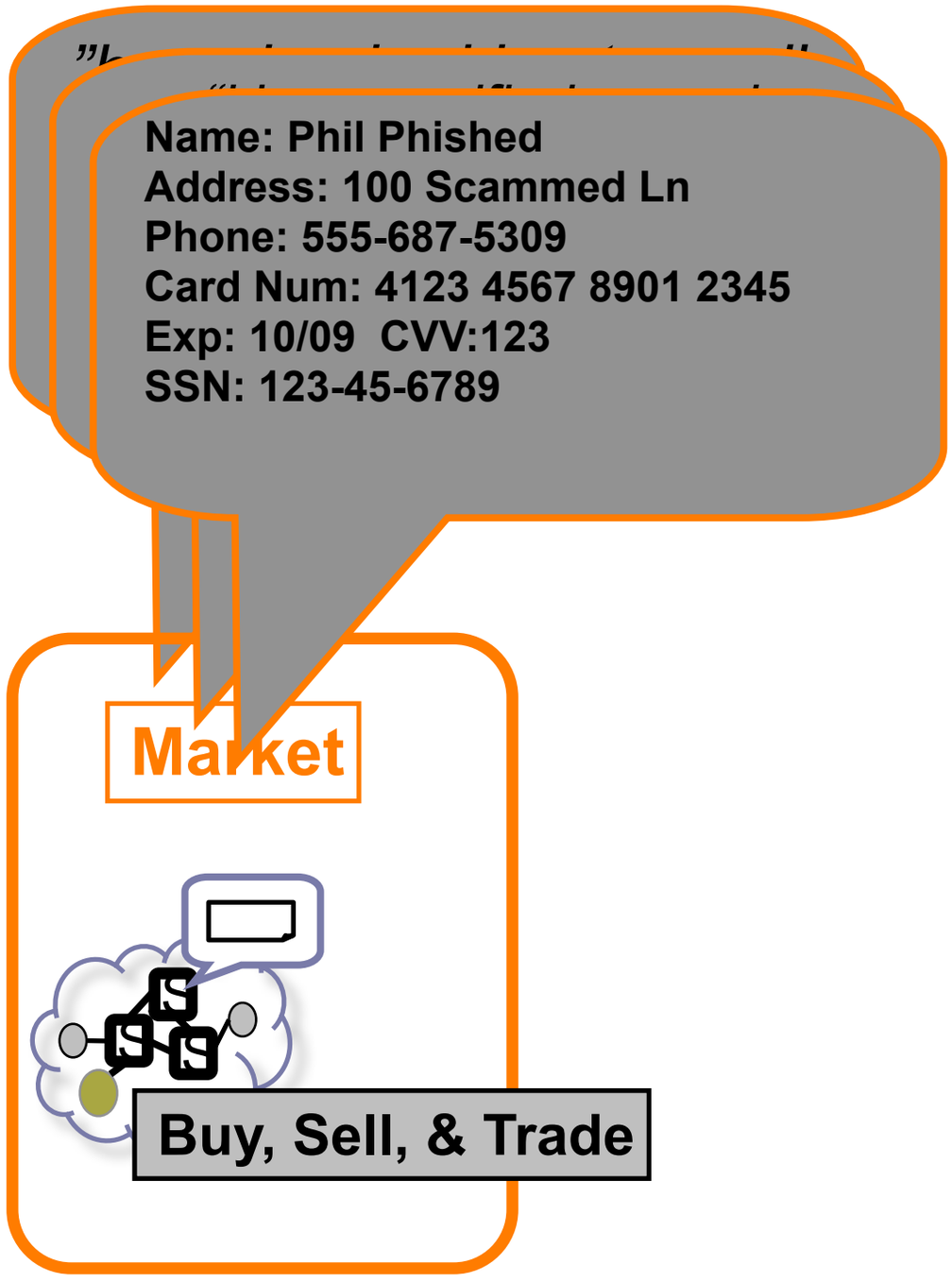
Market Organization and Data Collection

- Market is public channel active on independent IRC networks
- Common channel activity and admin. creates unified market
- IRC log dataset (2.4GB)
  - 13 million public messages
  - From Jan. ’06 to Aug. ‘06



Market Activity

- 1. Posting advertisements
  - Sales and want ads for goods and services
- 2. Posting sensitive personal information
  - Full personal information freely pasted to channel
  - Establishes credibility
- Need automatic techniques to identify ads and sensitive data



Measurement Methodology

- Three classes of measurement:
  - 1. Manual -> (Labeled dataset)
    - Manual classification of >3,500 messages with 60+ labels
      - Messages selected uniformly at random from corpus

Advertisement	Classification Label(s)
“have hacked hosts, mail lists, php mailer send to all inbox”	Hacked Host Sale Mailing List Sale Mailer Sale Ad
“i need 1 mastercard I give 1 linux hacked root”	Credit Card Want Hacked Host Sale

Measurement Methodology

- Three classes of measurement:
  - 2. Syntactic
    - Using regular expressions to pattern match structured sensitive data such as credit card numbers and SSNs

$cc = / \s\d{16}\s /;

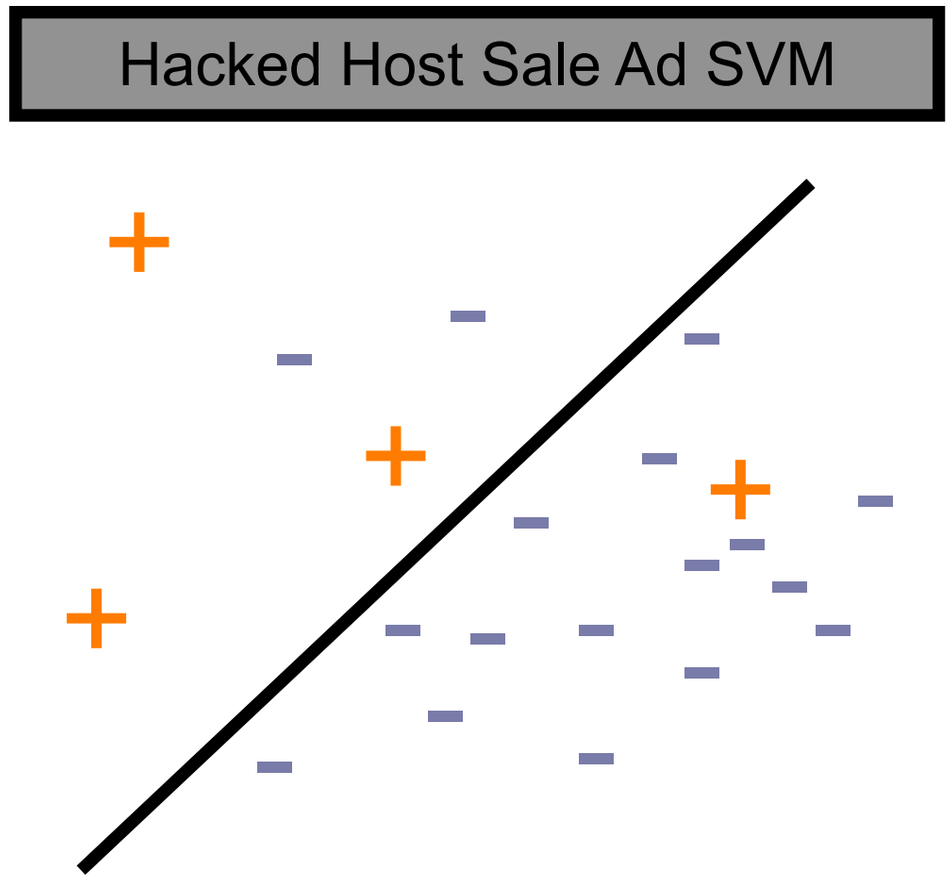
HaX0R: Free VISA! Name: Adrian Per… Num: 4123456789101234

HaX0R: SSN: 123-456-7859

Measurement Methodology

- Three classes of measurement:
  - 3. Semantic
    - Train binary SVM classifiers for each label using labeled dataset
    - Automatically classify messages

“have hacked hosts, mail lists, php mailer send to all inbox”



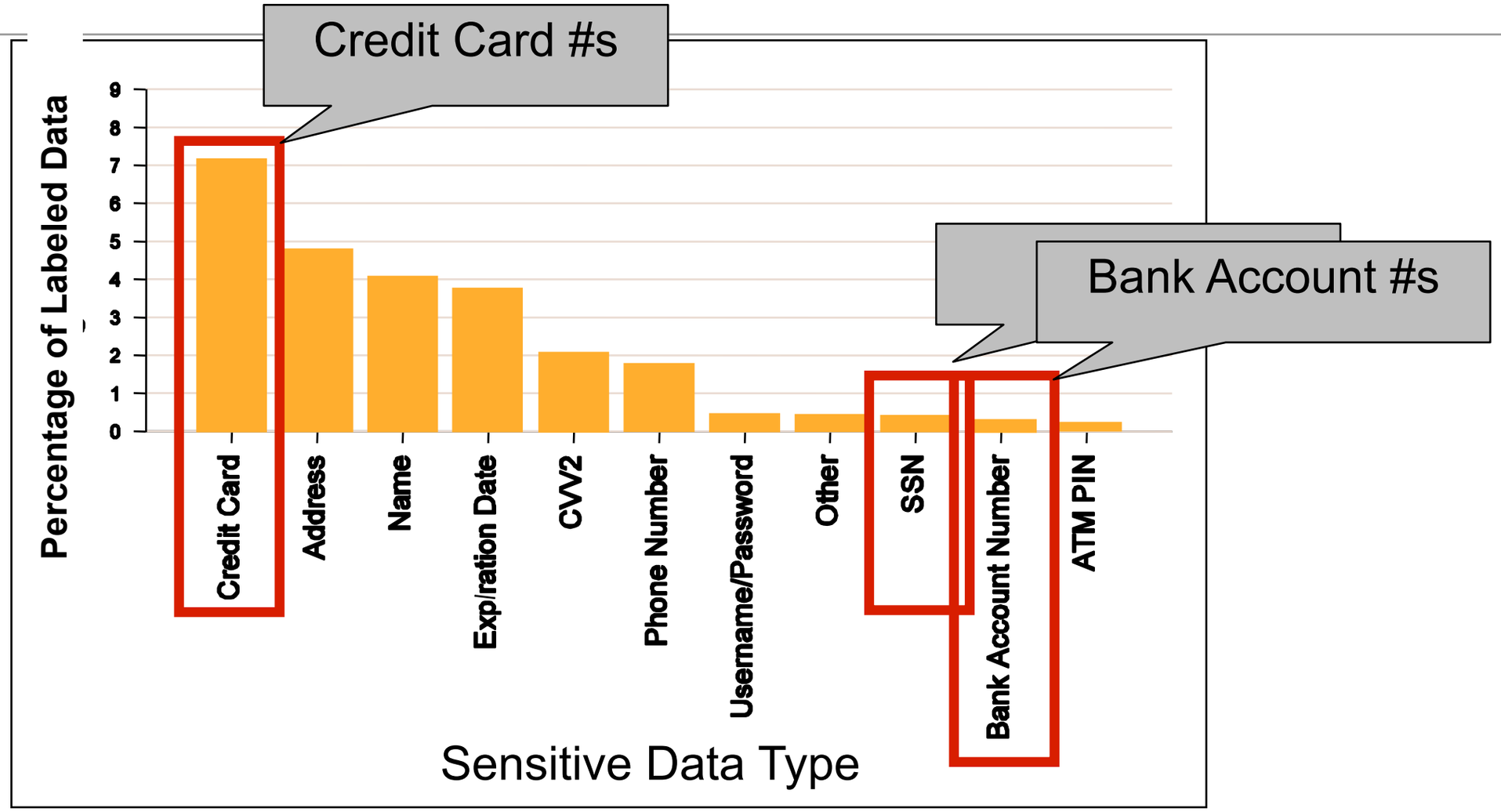
Measurement Complexities and Limitations

- No private messages
  - Limited transaction details and prices
- Assertions are not intentions
  - “Rippers” may advertise items they do not have
- Public market may bias behavior of miscreants
- Key Challenge: Validate data
  - Check Luhn digit, formats, valid ranges of SSNs
  - Cross-validate with other lists of compromised data
  - Need to collaborate with CC companies or law enforcement

Sensitive Data and Market Significance

- Is this market significant?
  - Measure sensitive data in corpus as indicator of significance
- Measurement Methodology:
  - Manually identify sensitive data in labeled dataset
  - Data validation
    - Checked that data was of valid format, in correct range
    - Verified Luhn digit on credit cards

Sensitive Data and Market Significance



- Measurement Results:
  - Credit cards compose 7% of labeled data
    - Estimate: 13 million line corpus * 7% = 910k (100k unique)
  - SSNs and bank accounts fall in 0.5 – 0.2% range

Estimating Wealth of Miscreants

- Goal: Estimate wealth stolen by market
- Measurement Methodology:
  - Transactions hidden by private channels
  - Median loss amount for credit/debit fraud = $427.501
  - Syntactic matches + Luhn check resulted in 87,143 potential cards
  - Include financial account data
- Measurement Results:
  - Credit card wealth = $37 million
  - Total: $93 million

Account Type	Total Balance
Balance	$18 million
Checking	$17 million
Mortgage	$15 million
Saving	$4 million

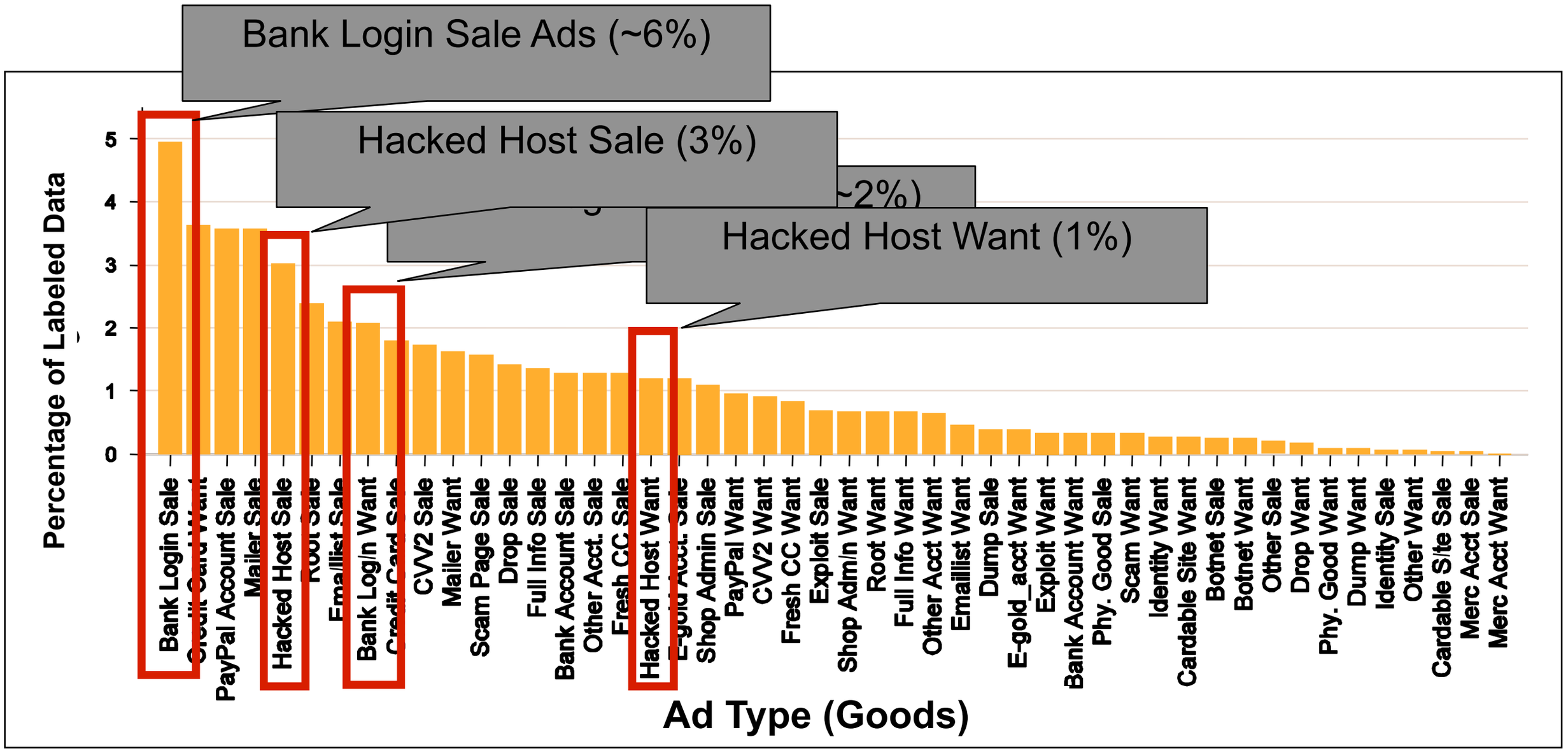
Table 1: Financial data totals from public posts.

*.2006 Internet Crime Complaint Center's Internet Crime Report

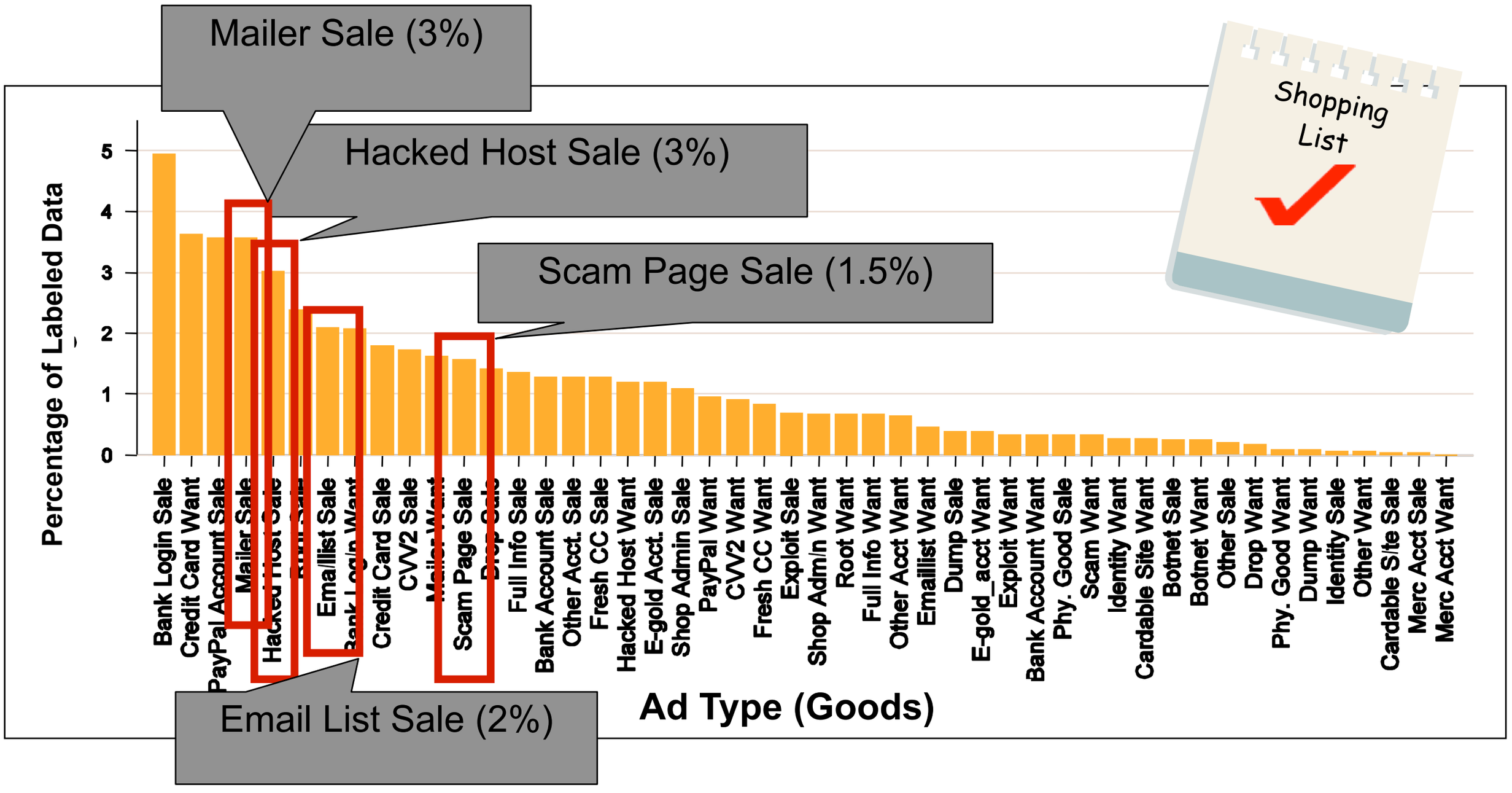
Goods, Services, and Prices

- Goods: Expansive collection of primarily virtual goods:
  - Online credentials and sensitive data
  - Hacking tools, spam kits, and phishing components
- Services: Fledgling service industry supports financial fraud:
  - Cashiers - Miscreant who converts data (credentials) to currency
  - Confirmers - Miscreant who poses as account owner/sender in money transfer
- Prices:
  - < 10% of advertisements include prices
  - Prices typically established in private messages

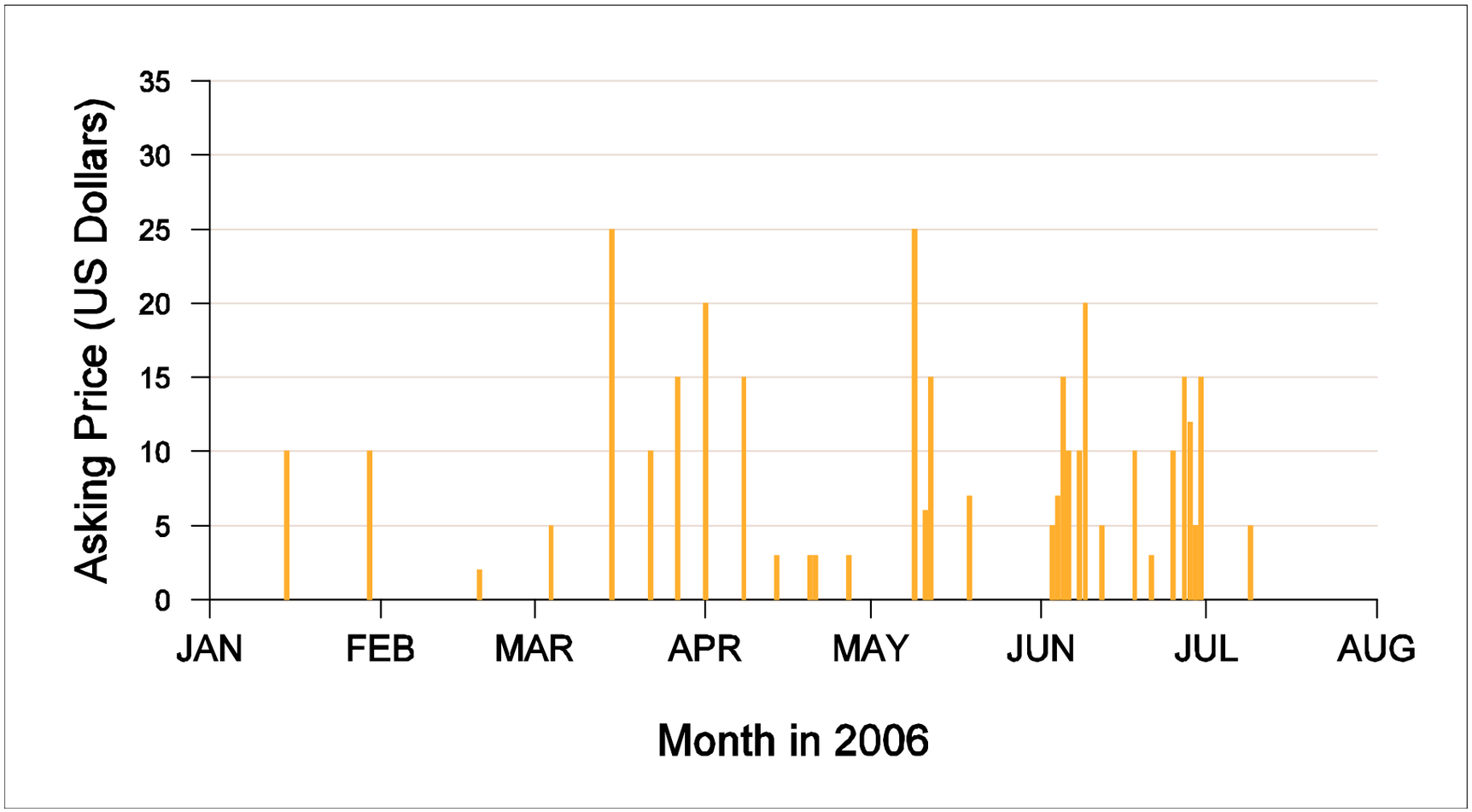
Distribution of Goods in Labeled Data



Distribution of Goods in Labeled Data



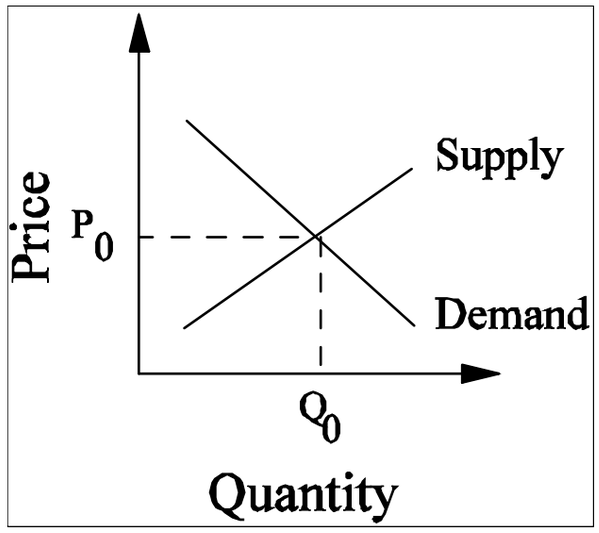
Asking Prices for Compromised Hosts



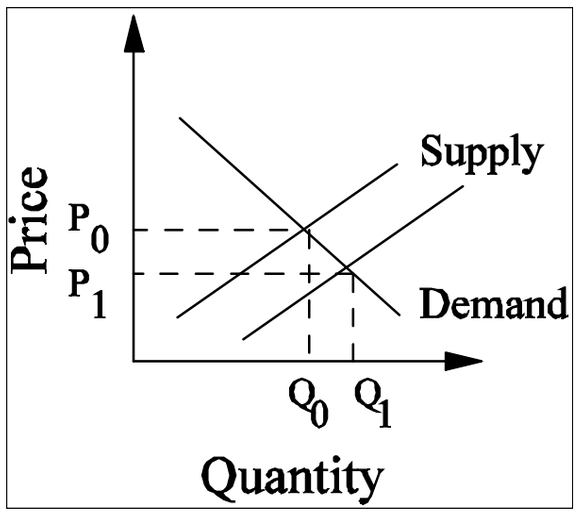
- Establishes cost to buy resources
  - May be useful to state strength of adversary in monetary terms
  - Cost to buy perhaps useful security metric?

World 0: A Wealth of Information

- Market may enable measurement of global trends and statistics
  - Idea: Price of a good in efficient market provides intersection of supply and demand curves
    - Assume a short-term constant demand
    - Then changes in price are result of shifts in supply curve
    - Increases or decreases in the quantity supplied



Supply and demand curves.



Shift of supply curve.

World 1: Markets Pose Security Threat

- Markets target of law enforcement activity:
  - U.S. Secret Service's Operation Firewall
    - July 2003 – late 2004, targeted administration
    - Required sting operation, inter-state, and multi-national cooperation
      - UK, Canada, Bulgaria, Belarus, Poland, Sweden, Netherlands, Ukraine
    - Resulted in arrest of 28, in 8 states, 6 countries
- Market reemerged after arrests
  - Decentralized, global nature of market makes traditional law enforcement activity time consuming and expensive
- Motivates need for more efficient low-cost countermeasures

Efficient Countermeasures

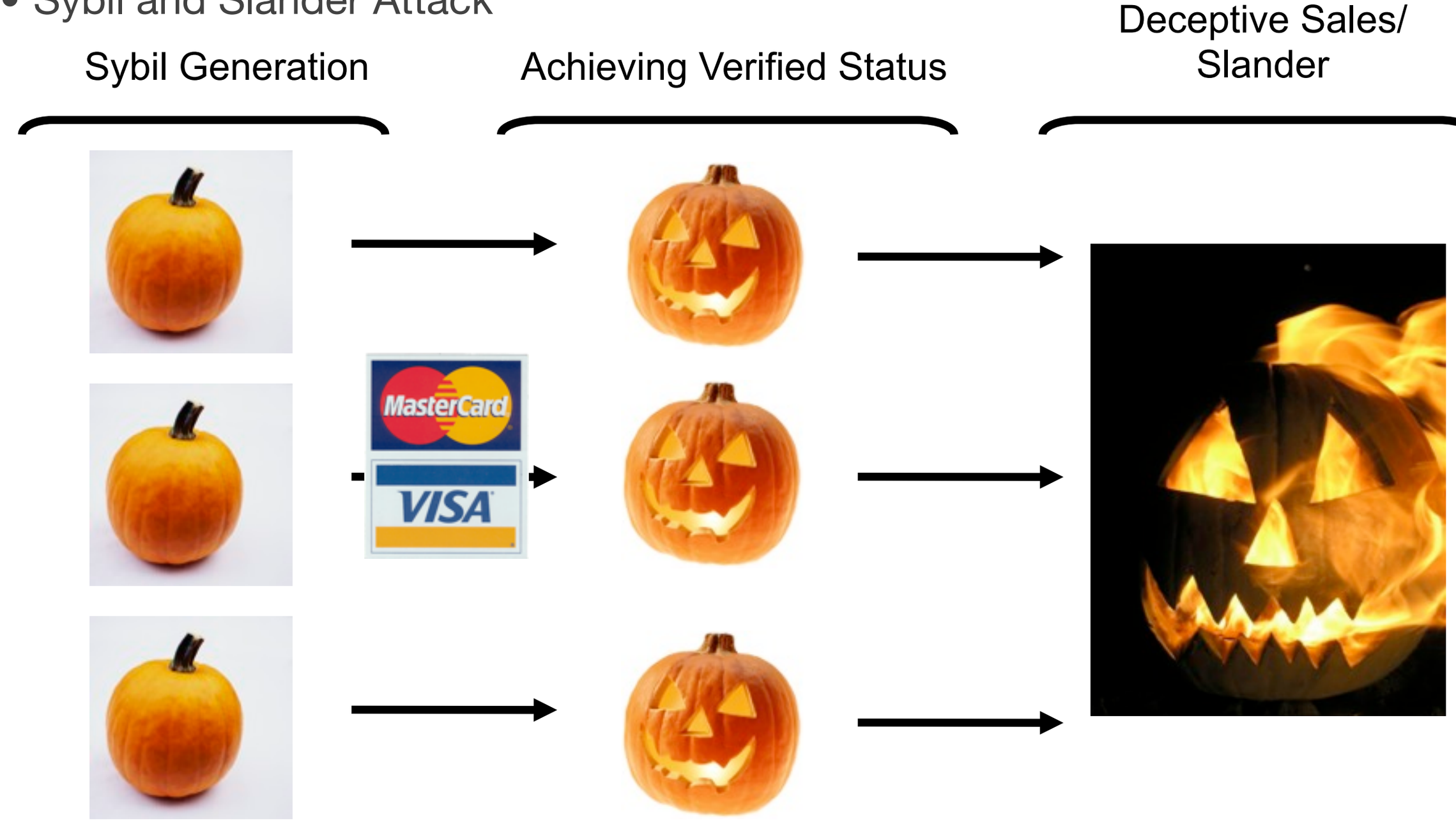
- Goal: Raise barrier to entry for eCrime
  - Reduce number of successful transactions
  - Push market towards closed market

- Approach: Establish environment which exhibits asymmetric information similar to Lemon Market
  - Buyers can’t distinguish quality of sellers

- Insight: Criminals will likely prefer anonymity over stronger verification system which relies on identity
  - Or we ease law enforcement’s job

Efficient Countermeasures

- Sybil and Slander Attack



Conclusion

- Shift from hacking “for fun” to “for profit” opens possible of modeling Internet-based crime as rational behavior (for profit)
- First study to systematically measure and analyze eCrime market
- Explored some beneficial uses of market-derived data & countermeasures
- Limitations of this study:
  - Soundness of measurement
    - Need for better verification and cross-validation
  - Completeness of measurement
    - What percentage of eCrime market activity are we seeing?
  - Applicability of measurements/conclusions
    - Can we apply our techniques to other eCrime markets?

Acknowledgments/References

- [Zhang] CANTINA: A Content-Based Approach to Detecting Phishing Web Sites was presented at, Yue Zhang, Jason I. Hong (presentation obtained from his website), and Lorrie F. Cranor, presented at www 2007.
- [Franklin] An Inquiry into the Nature and Causes of the Wealth of Internet Miscreants, Jason Franklin (presentation obtained from his website), Vern Paxon, Adrian Perrig, and Stefan Savage, presented at ACM CCS'07, Alexandria, VA, Nov. 2007.