SAND2013-3821C

CAN WE IDENTIFY SPEAR PHISHING TARGETS BEFORE THE EMAIL IS SENT?

Jeremy Wendt, JD Doak,

Andy Wilson, Roger Suppona

Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000..

MOTIVATION: SPEAR PHISHING

Attacker visits pages to find target data

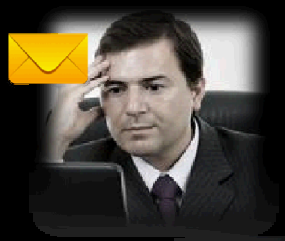
Logs record all visit data

Analysts identify and warn potential targets

Time Passes

Attacker crafts target-specific phishing email

Prepared user reports and deletes phishing email



Unprepared user follows link in email

Computer secure

Computer compromised

SOME IMPORTANT NOTES

- Spear Phishing moves at human speeds vs. network speeds
  - Attacker - identify targets, gather data, craft the email
    - hours to days
  - Victim - notice the email, read it, decide to click
    - seconds to hours/days
- Our Goal: Aid analysts in digesting logs

SCALE

- Here at Sandia, we have a “medium sized” web-presence
  - 55 different machines serve web pages
  - 254 domain names served
  - ~2.5M distinct URLs found on SNL servers by crawling
  - 500K-1M entries per day
  - ~36K unique URLs requested per day
  - ~15K unique visitors per day
  - Downloading a single web page creates 1-20 entries in the log
    - HTML, images, CSS, JavaScript, etc.
- And portions of the data are likely to be false

DATA

- Each entry contains many pieces of data
  - Timestamp
  - Client IP
  - Client user agent string (UAS)
  - Requested URL
  - Refer string
  - X-Forwarded For (XFF)
  - Much more

HELP THE ANALYST

- How can an analyst sift through such a mass of data quickly enough to find actionable data?
  - Improve the data, filter the data, sort the data, and present the data for better analyst triage
    - Distinguish crawler traffic from browser traffic
    - Sort the results so that most “interesting” sits at the top
    - Display the data so that a big picture is quickly visible

DISTINGUISH CRAWLER FROM BROWSER

- Goal: Separate crawler traffic (e.g., Google indexer) from human-driven browser traffic
  - Some interesting other “patterns” could be hidden by bot “noise”
- UAS can be used to identify bot or not
  - But UAS is client-provided and can be falsified (or left empty)
  - Use UAS as initial grouping, look for other characteristics that distinguish the groups
    - Remember: Some of the clients are almost certainly lying
- NOTE: Visitors must leave 20 entries in the log to be included in this analysis
  - Really ~3-5 distinct webpages

BOT CHARACTERISTICS

- Regular and fast: Get as many pages as quickly as you can!
- (Often) Polite: Don't hit a server too often or you'll kill it (and get blacklisted)
- Busy: Crawling the whole Internet takes a while
- (Often) Distributed: Only one machine needs to download each page
- Long memories: Don't redownload a resource until TIMEOUT passes

BROWSER CHARACTERISTICS

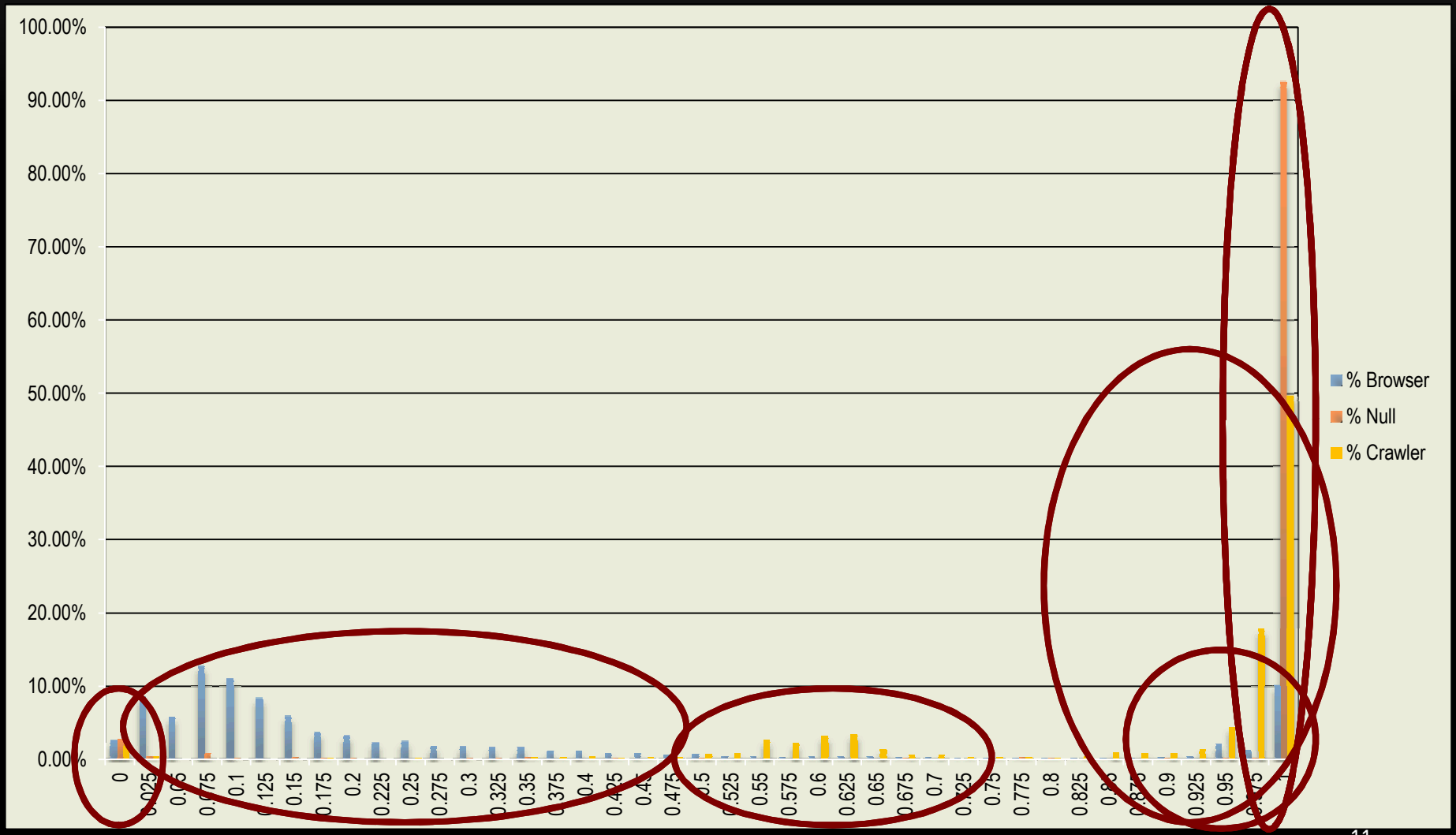
- Bursty: Download all images/css/js/etc. for a page NOW!
  - Then do nothing until the user clicks again
- Lots of different file types
  - HTML/images/css/js/etc.
- Shorter memory: Will redownload content more regularly

PERCENT HTML

- Browsers pull down 3-20 non-HTML documents for each HTML document to render the webpage
- Bots care most about indexing text (so often don't need images, etc.), and cache supporting documents longer than browsers

- %HTML = NumHTML / NumTotal

PERCENT HTML

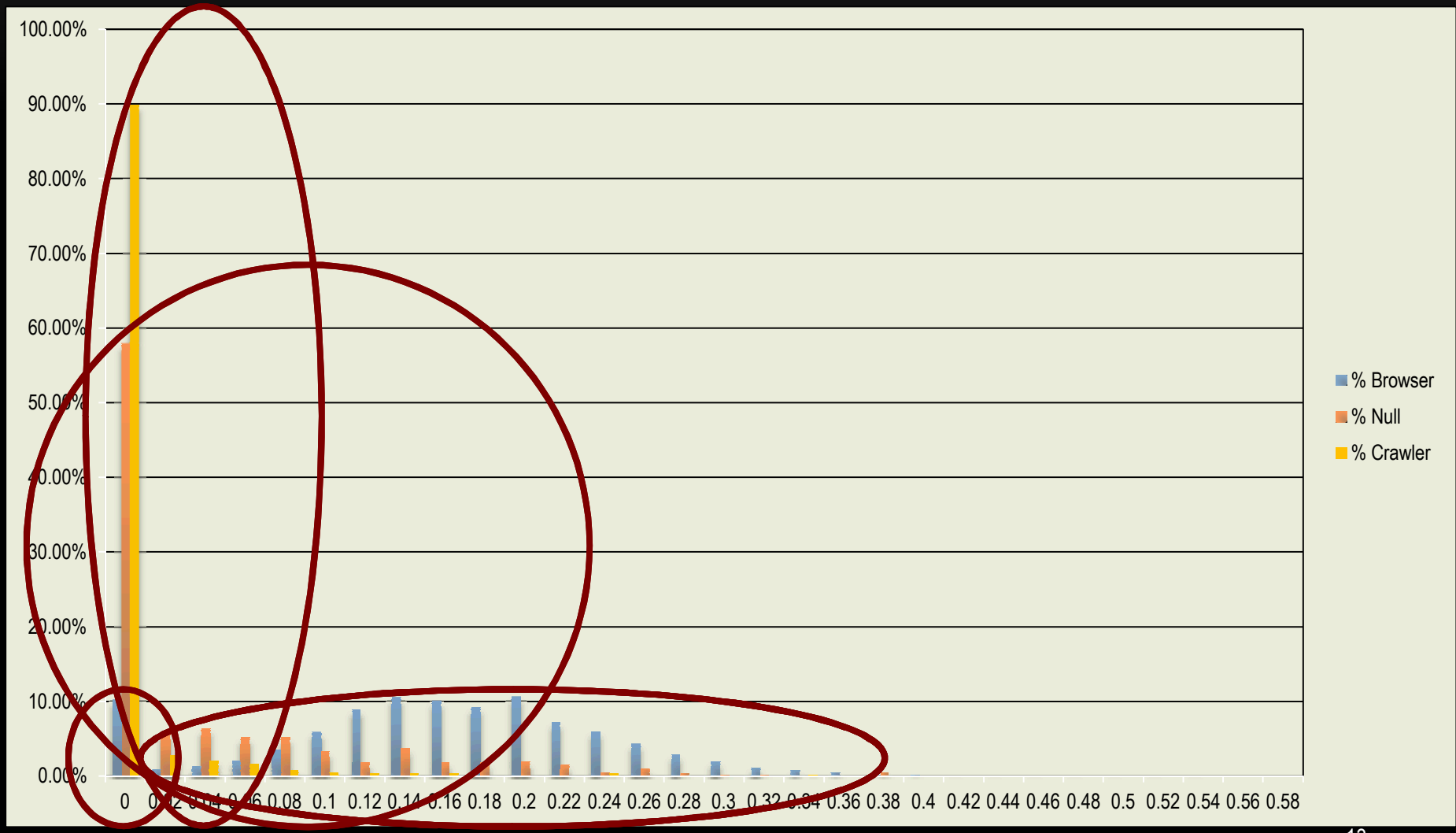


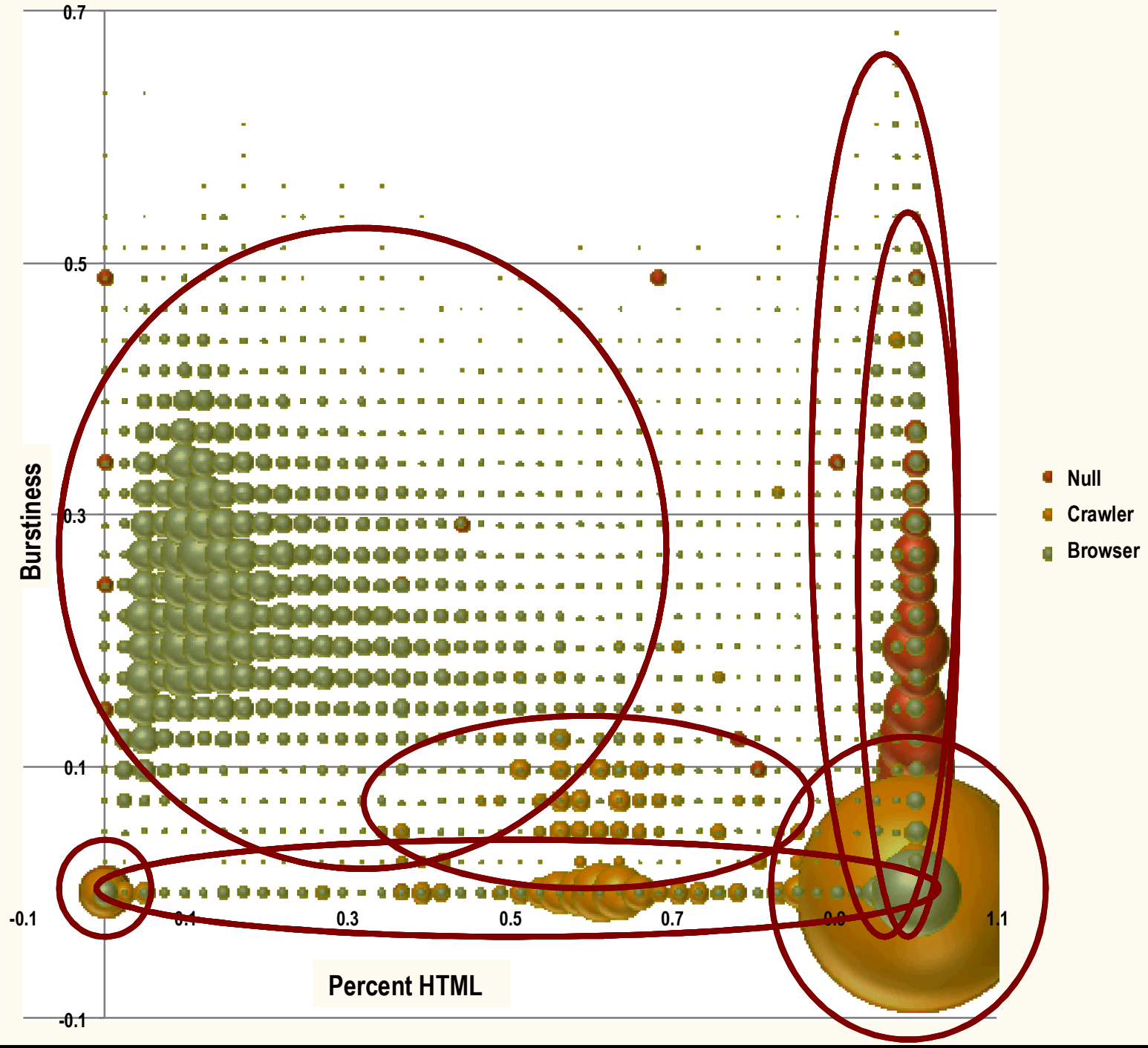
BURSTS-TO-VISITS RATIO

- After browsers retrieve the HTML, they quickly parse it and request all supporting documents
- Bots only request pages from a site every second or so to keep from being blacklisted

- A “burst” is defined as more than N visits in an M second window by the same visitor
- Ratio = NumBursts / NumVisits

BURST-TO-VISITS RATIO




0.7
0.5
0.3
0.1
-0.1
Burstiness
-0.1
0.1
0.3
0.5
0.7
0.9
1.1
Percent HTML
Null
Crawler
Browser

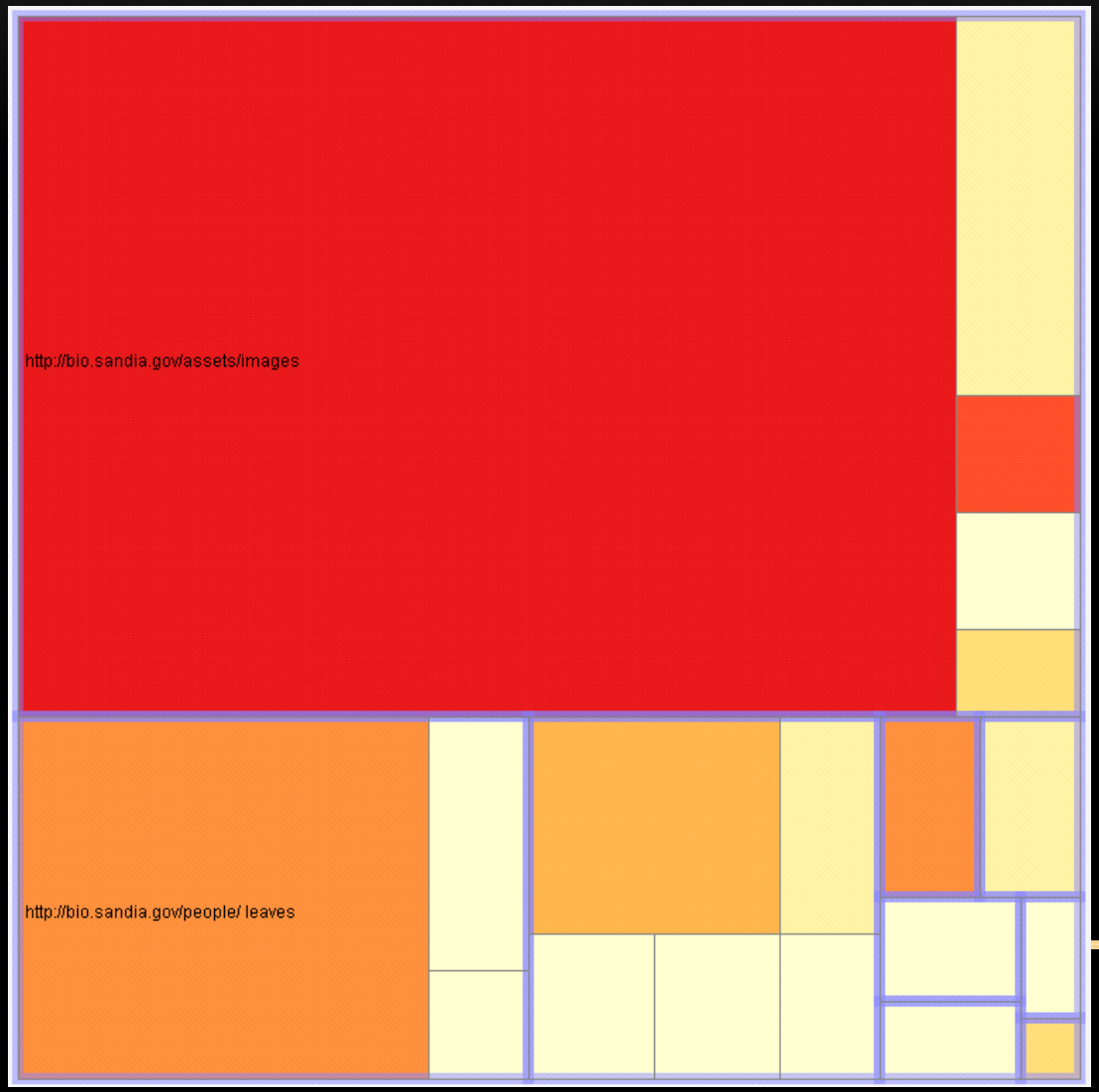
BOT-OR-NOT RESULTS

- Accuracy after combining the two features
  - UAS says “bot”: 98.6%
  - “null” UAS (assumed bot): 96.8%
    - NOTE: Evidence of ~1% browser-based null UAS
  - UAS says “browser”: 81.8%
    - NOTE: Evidence of 10-17% is bot

BOT-OR-NOT RESULTS

- Visitors
  - 10.3K identified bots
  - 38.6K identified browsers
  - 166K too few visits
- Visits
  - 3.2M made by bots
  - 3M made by browsers
  - 0.8M made by too few visits
- Liars (possibly)
  - 8.7K say they’re browser, but aren’t
  - 30 (yes, thirty) say they’re bot or null, but aren’t

TREEMAP



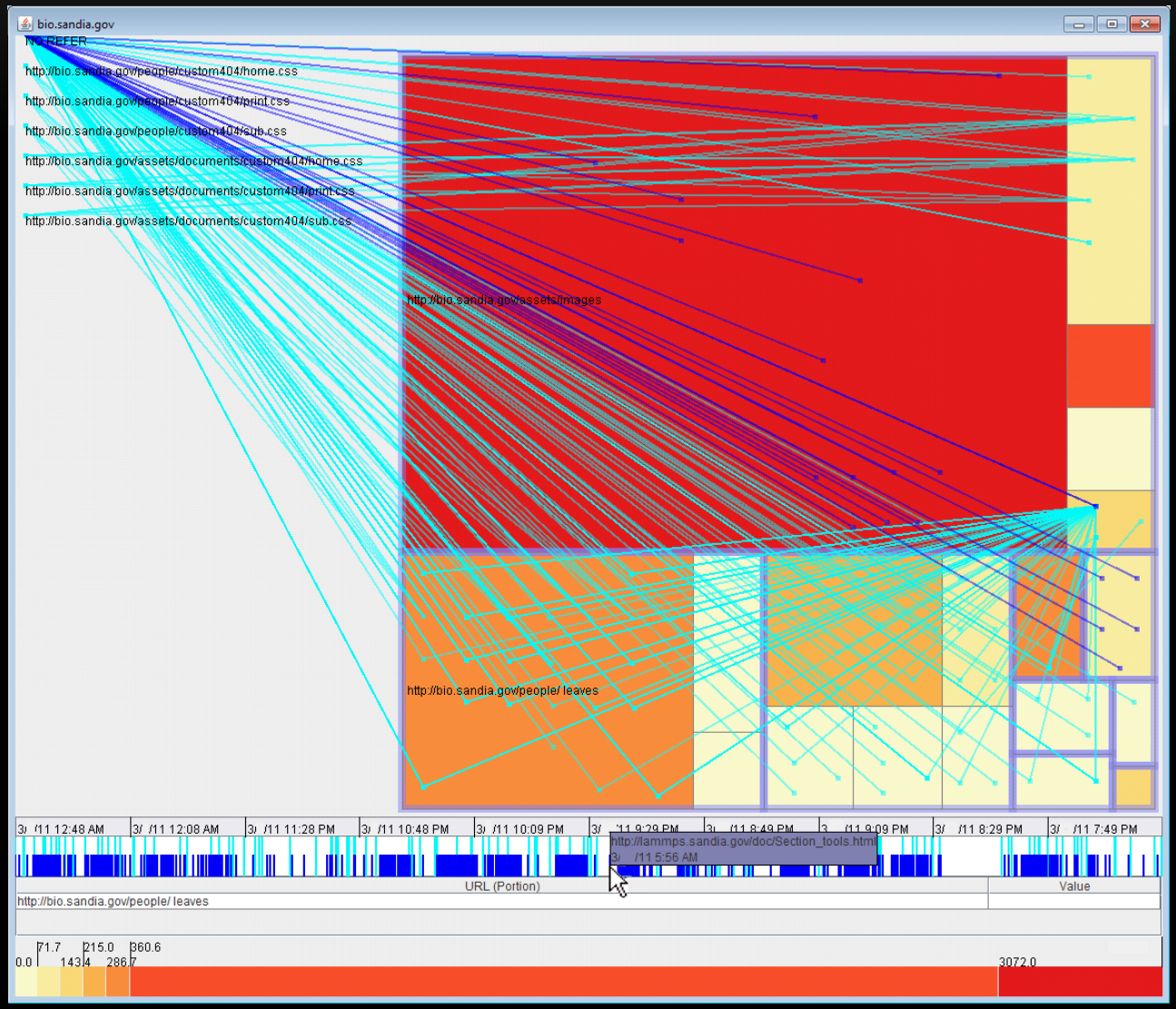
A rectangle's area represents the number of distinct webpages within each subdirectory.

A rectangle's color indicates the number of visits to webpages within each subdirectory.

The blue-outlined rectangles represent the level-one directories on bio.sandia.gov.

The grey-outlined rectangles represent the level-two directories on bio.sandia.gov.

TRAVEL BY MAP



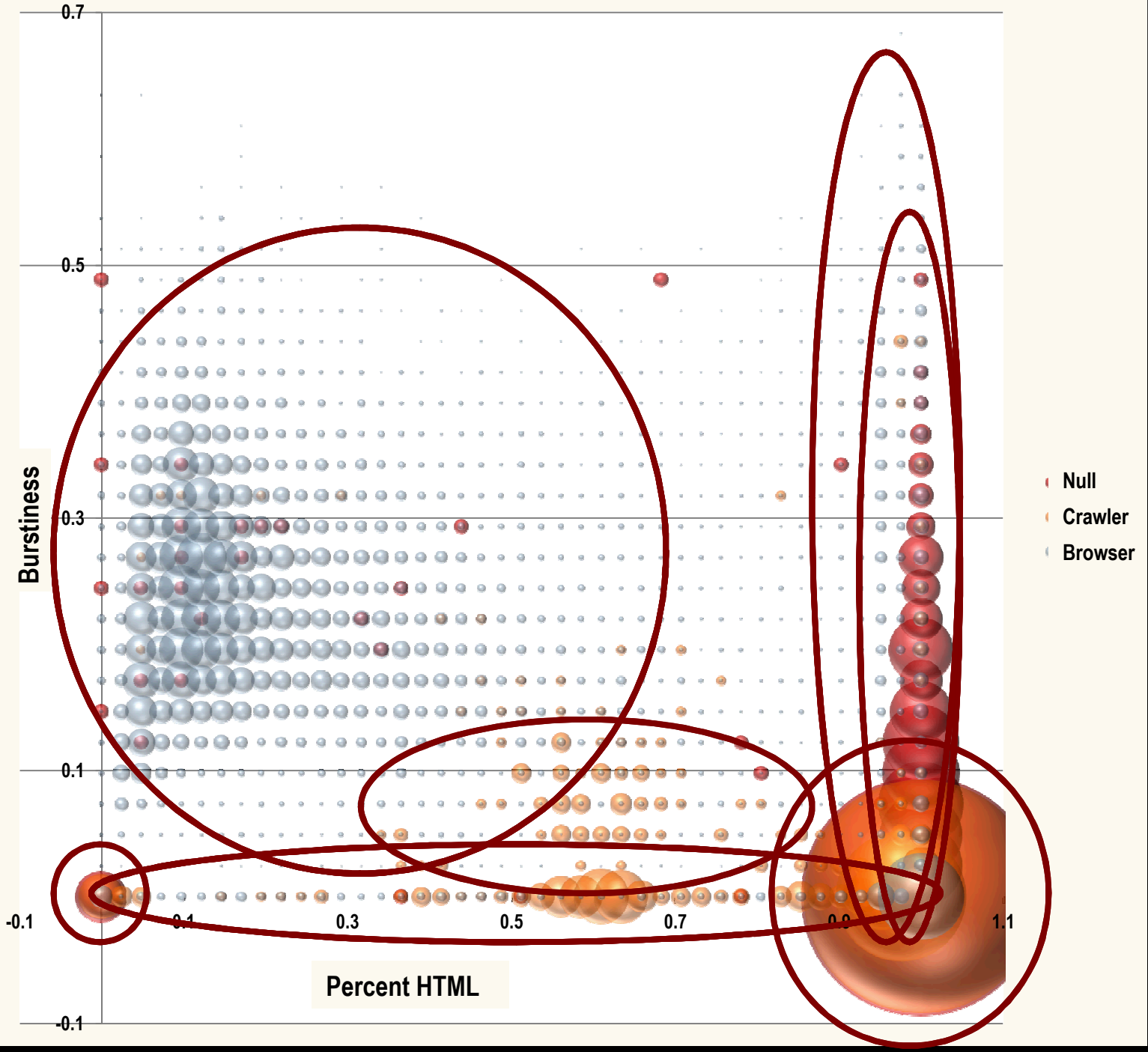
QUESTIONS?

BACKUP SLIDES

CLIENT'S UNIQUE IDENTIFIER

- Due to DHCP, the same IP address can be used by different machines at different times
- Different applications on the same machine make HTTP requests for different reasons
  - UAS
- Proxied HTTP requests often use the last proxy server's IP address as the "client IP"
  - XFF (can contain private-space IPs which aren't globally unique)

- We decided to use the following pattern
  - "IP"/UAS
    - "IP" – If no XFF, use IP, else if 1st XFF addr is unique, use it, else use 1st XFF addr/1st unique address in XFF


0.7
0.5
0.3
0.1
-0.1
Burstiness
-0.1
0.1
0.3
0.5
0.7
1.1
Percent HTML
Null
Crawler
Browser

TRIAGE SUPPORT