I never signed up for this!

Privacy implications of email tracking

Steven Englehardt
@s_englehardt
senglehardt.com

Joint work with:
Jeffrey Han and
Arvind Narayanan

This research is funded by NSF grant CNS 1526353, by a research grant from Mozilla, and by an Amazon AWS Cloud Credits for Research grant.
On Sunday, August 27, 2017 1:19 PM, Bed Bath & Beyond <BedBath&Beyond@emailbedbathandbeyond.com> wrote:

```
Wamsutta® sheets clearance. Shop now! View as web page

Bed Bath & Beyond®

FREE SHIPPING ON ORDERS OVER $29
Get Details

IN-STORE PICKUP
RESERVE ONLINE PAY IN STORE Learn More

DON'T FORGET IF YOU HAVEN'T ALREADY REDEEMED! 20% OFF ONE SINGLE ITEM IN-STORE OR ONLINE EXCLUSIVE OFFER FOR THIS EMAIL ADDRESS ONLY SHOP NOW

OFFER EXPIRES 9/10/17 VIEW DETAILS

UP TO 60% OFF SELECT WAMSUTTA® 525 & 620 THREAD COUNT SHEETS SNAG GREAT PRICES ON TWIN/TWIN XL SHEETS JUST IN TIME FOR COLLEGE! SHOP NOW! WAMSUTTA® SINCE 1846
```
Who doesn’t love reading email like this?
What are the privacy implications?

Remote content enabled
Email Tracking

Know the second a lead opens an email, send a perfectly timed follow-up, and close deals faster than ever.

<table>
<thead>
<tr>
<th>Priority</th>
<th>Name</th>
<th>Email Message</th>
<th>Opened Times</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adam Carpenter</td>
<td>Your Biglytics Trial</td>
<td>2 opens</td>
</tr>
<tr>
<td>Today</td>
<td>Adam Carpenter</td>
<td>Your Biglytics Trial</td>
<td>2 opens</td>
</tr>
<tr>
<td>Yesterday</td>
<td>Someone</td>
<td>The Future of Big Data</td>
<td>LIVE Webinar</td>
</tr>
</tbody>
</table>
Emails are tracked far beyond send tracking
Your device contacts 24 companies:
→ 20 can track you (if supported)
→ 10 receive an MD5 hash of your email address

### Receives MD5(email address) & Sets a Cookie
- American List Counsel (alcmpn.com)
- LiveIntent (liadm.com)
- Oracle (nexac.com)
- Acxiom (rlcdn.com, pippio.com, acxiom-online.com)
- Criteo (criteo.com)
- Conversant Media (dotomi.com)
- V12 Data (v12group.com)
- VideoAmp (videoamp.com)
- <Unknown> (alocdn.com)

### Receives MD5(email addr.)
- Criteo (emailretargeting.com)
- Neustar (agkn.com)

### Receives Bare Request
- LiveIntent (licasd.com)
- Google (2mdn.net)
- Akamai (akamai.net)

### Sets a Cookie
- OpenX (openx.net)
- comScore (scorecardresearch.com, voicefive.com)
- Oracle (bluekai.com)
- Google (doubleclick.net)
- Realtime Targeting Aps (mojn.com)

### MediaMath (mathtag.com)
- TapAd (tapad.com)
- IPONWEB (bidswitch.net)
- AOL (advertising.com)
- Centro (sitescout.com)
- The Trade Desk (adsrvr.org)
- Adobe (demdex.net)
Email Tracking ≈ Web Tracking - Javascript
Measuring email tracking at scale

1. Crawled 15,700 sites
2. Signed up for mailing lists
3. Received 13,000 emails from ~900 sites
4. Measured tracking with OpenWPM
Our Findings
Many of the top web trackers are in emails

<table>
<thead>
<tr>
<th>Domain</th>
<th>% of Emails</th>
<th>% of Top 1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>doubleclick.net</td>
<td>22.2</td>
<td>47.5</td>
</tr>
<tr>
<td>mathtag.com</td>
<td>14.2</td>
<td>7.9</td>
</tr>
<tr>
<td>dotomi.com</td>
<td>12.7</td>
<td>3.5</td>
</tr>
<tr>
<td>adnxs.com</td>
<td>12.2</td>
<td>13.2</td>
</tr>
<tr>
<td>tapad.com</td>
<td>11.0</td>
<td>2.6</td>
</tr>
<tr>
<td>liadm.com</td>
<td>11.0</td>
<td>0.4</td>
</tr>
<tr>
<td>returnpath.net</td>
<td>11.0</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>bidswitch.net</td>
<td>10.5</td>
<td>4.9</td>
</tr>
<tr>
<td>fonts.googleapis.com</td>
<td>10.2</td>
<td>39.4</td>
</tr>
<tr>
<td>list-manage.com</td>
<td>10.1</td>
<td>&lt;0.1</td>
</tr>
</tbody>
</table>

85% of emails embed third parties (with an average of 5 per email)
29% of emails (from 19% of senders) leak the email address to third parties

<table>
<thead>
<tr>
<th>Leak</th>
<th># of Senders</th>
<th># of Recipients</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD5</td>
<td>100</td>
<td>38</td>
</tr>
<tr>
<td>SHA1</td>
<td>64</td>
<td>19</td>
</tr>
<tr>
<td>SHA256</td>
<td>69</td>
<td>13</td>
</tr>
<tr>
<td>Plaintext Domain</td>
<td>55</td>
<td>2</td>
</tr>
<tr>
<td>Plaintext Address</td>
<td>77</td>
<td>54</td>
</tr>
<tr>
<td>URL Encoded Address</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>SHA1 of MD5*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SHA256 of MD5*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MD5 of MD5*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SHA384</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Trackers can correlate email and web tracking

1. Save cookie and email
2. Check database with cookie
3. Use stored email to track visits and target ads
“People-based” Marketing
As an identifier, email is both deterministic and persistent. That is, when a consumer gives out a verified email, it usually belongs to only that consumer. That can’t be said of all typical advertising identifiers. Cookies, for example, live on desktop browsers that are often shared with no way to distinguish who’s using it. And whereas email is cross-device, cookies aren’t.

LiveIntent may also receive non-personal information from online and offline sources, including the types described below, from our business partners.
To de-identify this information, either we or our business partners [hash it].

we use a double hashing method ... to ensure the non-reversibility of your information. A hash of your email corresponds to a series of characters that does not permit your identification.
Does hashing protect user privacy?

Email Hash
b5184f3fb0fe35e4319b729f05017f6e

Tracker Database

Tracking Data
- https://www.webmd.com/cancer/default.htm
- http://www.foxnews.com/
- Living social Healthy Living email campaign
- $105 in Personal Health purchases from CVS
- $55 purchase from Babies’R’Us
Does hashing protect user privacy?

Tracker Database

Email Hash
b5184f3fb0fe35e4319b729f05017f6e

Tracking Data
- https://www.webmd.com/cancer/default.htm
- http://www.foxnews.com/
- LivingSocial Healthy Living email campaign
- $105 in Personal Health purchases from CVS
- $55 purchase from Babies”R”Us

Run a “re-identification attack” yourself! Open your terminal and enter:

Linux: echo -n ste@cs.princeton.edu | md5sum

MacOS: echo -n ste@cs.princeton.edu | md5
Easy  ste@cs.princeton.edu  →  b5184f3fb0fe35e4319b729f05017f6e

Hard  b5184f3fb0fe35e4319b729f05017f6e  →  ste@cs.princeton.edu
Easy

ste@cs.princeton.edu  →  b5184f3fb0fe35e4319b729f05017f6e

Hard

b5184f3fb0fe35e4319b729f05017f6e  →  ste@cs.princeton.edu

Easy (when you can guess the possible inputs)
jh34@alumni.princeton.edu  →  261495fd24d108b3c573527b3854af00
ste@cs.princeton.edu  →  b5184f3fb0fe35e4319b729f05017f6e
arvindn@cs.princeton.edu  →  16eaf6d2cef77e145db18804d2aa4fd5
Email addresses aren’t secrets!

Use email database leaks...

...and just guess the rest.

GPU cloud computer: $24.48 / hour → 450 billion MD5 hashes / second

~4.7 billion email addresses total. If we generate a real address every 1 in 1 million guesses, we can generate the entire space for less than $75.

Past research recovered 45-70% of emails.

More info:

Don’t want to guess? Reverse hashes for $0.04/email

- theleadswarehouse.com
- infutor.com
- datafinder.com
Takeaways

1. The line between email and web tracking is blurry
2. Email addresses are commonly leaked to trackers in emails
3. Claims of “de-identification” are suspect

More Info

- **Full paper:** [https://senglehardt.com/papers/pets18_email_tracking.pdf](https://senglehardt.com/papers/pets18_email_tracking.pdf)
- **More on identity leaks:** [https://freedom-to-tinker.com/tag/noboundaries/](https://freedom-to-tinker.com/tag/noboundaries/)