Abuse at scale

hearn@google.com

Agenda

Agenda

1. Stories from abuse@google.com
2. Abuse in 2012
3. Abuse report handling
    a. Why it's hard
    b. What we could do about it

Stories from the abyss

Gmail then

Launched 2004, invite only. 2006, open invites.

- Gmail does not provide sender IP for web sends
- Open signups make abuse fighting much harder
- CAPTCHA solving teams became available, $1 per thousand CAPTCHAs.
- **Result**>50% of all outbound mail is spam within months

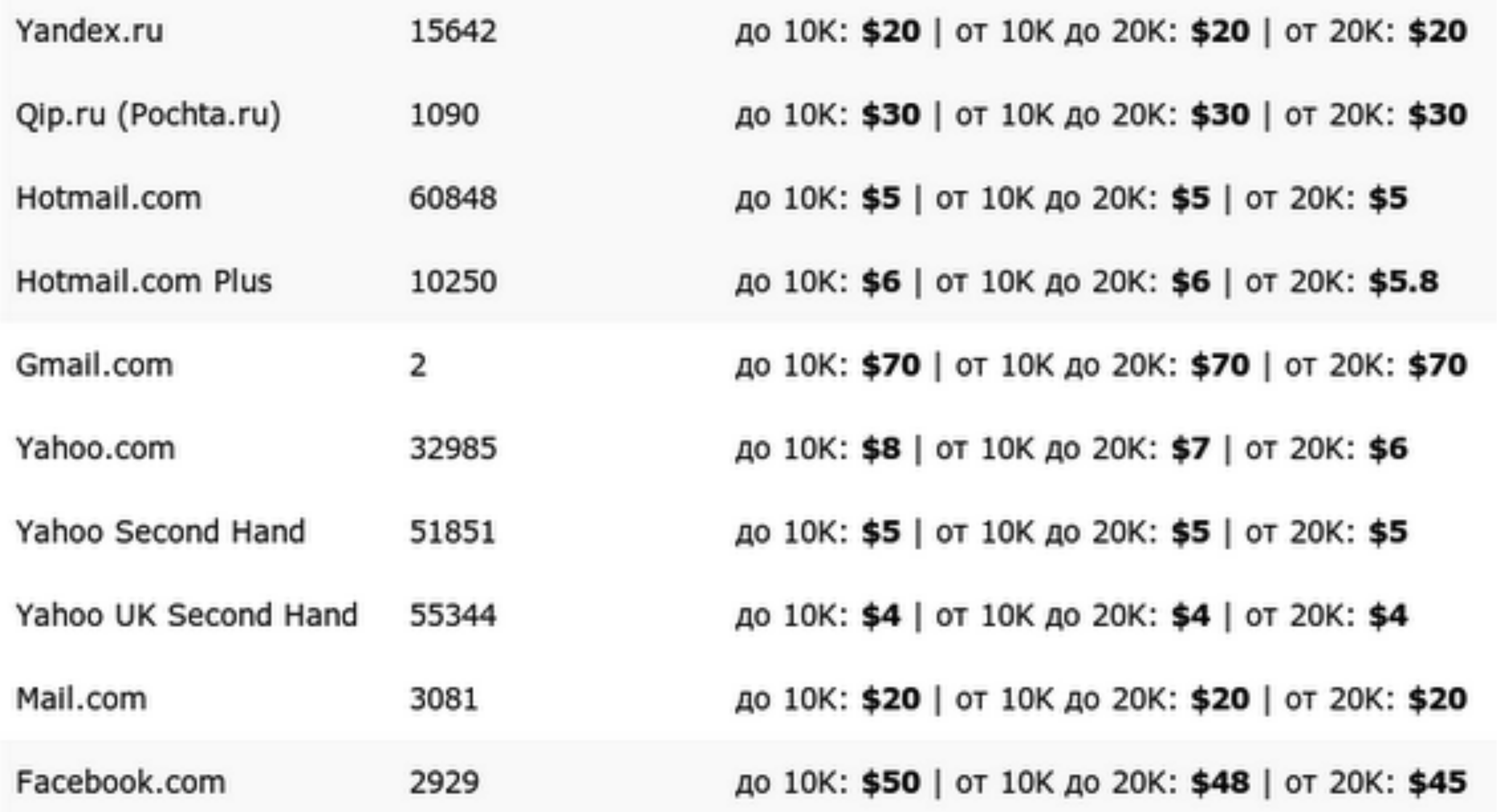
Gmail abuse team split out from inbound spam and grown

Gmail now

- No major outbound campaigns using spammy accounts
- Disclaimer: still send 5,000 (legit) mails/sec
    - you may get sometimes get mail from @gmail.com accounts that you don't want

How?

- Mail send risk analysis with hundreds of features, ML
- Phone verification on suspect spamming accounts
- Tactical operations against account sellers
- Account signup protected by risk analysis/ML/encrypted javascript, dedicated team that monitors bulk signup



Yandex.ru	15642	до 10K: **$20** | от 10K до 20K: **$20** | от 20K: **$20**
Qip.ru (Pochta.ru)	1090	до 10K: **$30** | от 10K до 20K: **$30** | от 20K: **$30**
Hotmail.com	60848	до 10K: **$5** | от 10K до 20K: **$5** | от 20K: **$5**
Hotmail.com Plus	10250	до 10K: **$6** | от 10K до 20K: **$6** | от 20K: **$5.8**
Gmail.com	2	до 10K: **$70** | от 10K до 20K: **$70** | от 20K: **$70**
Yahoo.com	32985	до 10K: **$8** | от 10K до 20K: **$7** | от 20K: **$6**
Yahoo Second Hand	51851	до 10K: **$5** | от 10K до 20K: **$5** | от 20K: **$5**
Yahoo UK Second Hand	55344	до 10K: **$4** | от 10K до 20K: **$4** | от 20K: **$4**
Mail.com	3081	до 10K: **$20** | от 10K до 20K: **$20** | от 20K: **$20**
Facebook.com	2929	до 10K: **$50** | от 10K до 20K: **$48** | от 20K: **$45**

Account sellers still exist. Normal price is $120-$150 per thousand (phone verified)

This price level makes bulk spam uneconomic.

Problem areas

- Spammers who pay for the ability to spam
- Spammers who claim they will pay but don't
- 10,000+ engineers/product managers who are not used to thinking adversarially
- Highly motivated spammers who find exploits
  - Students love Gmail. Let's make it available to universities!
  - Spammer discovers he can make fake universities: *.edu.tk is treated as valid (now fixed)
  - CAPTCHAs that are open to replay attacks
  - .... etc

Google abuse in 2012

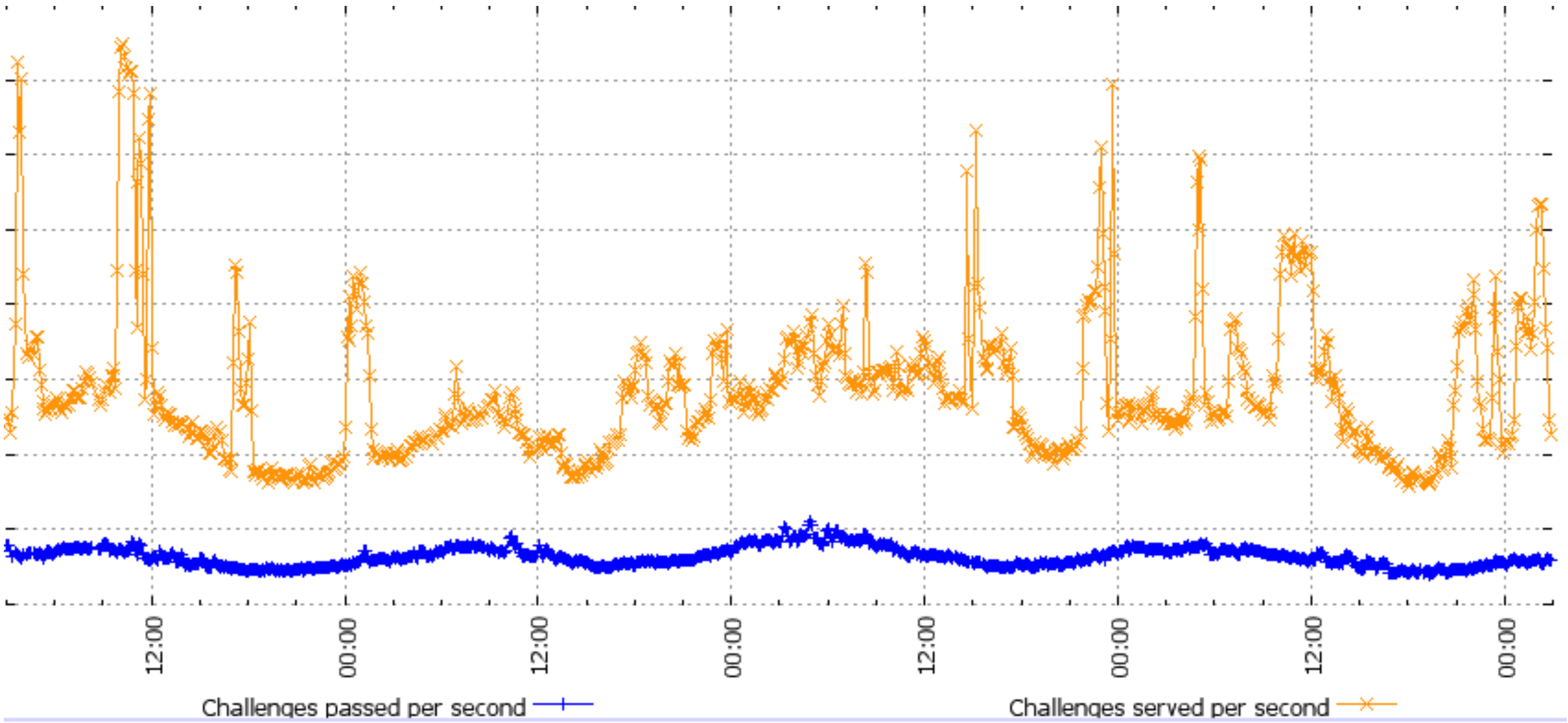
Recent trends

April 2010 - the world changed

- Bulk signup era is over
- Account hijacking begins
    - Over 1 million sets of credentials tried per day
    - Successfully authenticating to >100,000 accounts per day

WTF?

The age of the password is **over and never coming back**


12:00
00:00
12:00
00:00
12:00
00:00
12:00
00:00
Challenges passed per second
Challenges served per second

Solution

Abuse team becomes anti-hijacking team

Online login risk analysis

- Classifies 60-100k logins per second (2-3k/sec web)
- <100msec
- 0.1% false positive rate

2 years later, web hijacking on Gmail is largely wiped out.

Abuse report handling

Nobody expects the Spanish Inquisition!

abuse@gmail.com

Some unhappy truths:

- Receives >40 reports/second
- Reports grouped into "feeds"
- Automatically reviewed in almost all cases
- Abuse report handling is **a hard problem**

Why is processing hard?

- Finding trusted feeds is tricky
  - Individual reports have wildly varying quality, useful only in aggregate
  - "Trusted partners" are incentivized to become untrusted partners
  - Abuse reporting mechanisms frequently gamed
- Trustworthiness is not enough. You have to add coverage too.
  - If you have <100 users it makes no difference.
  - Abuse feed agreements exist between most major players, hard to avoid spamming them

Why is sending hard?

- Abuse reports contain verbatim/lightly redacted copies of mails
- Users have an expectation of privacy
- People click "report spam" on mails which are not spam
- Receivers should be processing abuse reports *from* us automatically and with reasonably good privacy controls:
  - Manual review for sanity checking: OK
  - Manual review of most abuse reports: NOT OK

What works best?

- Feeds that aggregate large numbers of users
- Feeds that have active anti-abuse teams behind them
    - Otherwise spammers will game the system
- Feeds that use standard formats like ARF
- Feeds which are automated

Ideas for moving forward

- Upgrades to ARF:
  - Could distinguish "this is spam" from "this is from a friend but doesn't seem like them".
    Easy extension to Feedback-Type.
  - URL abuse (goo.gl)
- Self-service tool for @google abuse feeds?

- Neutral / non profit aggregators that enforce basic ground rules?

The end!

Thanks for listening