Reverse Bayesian poisoning:

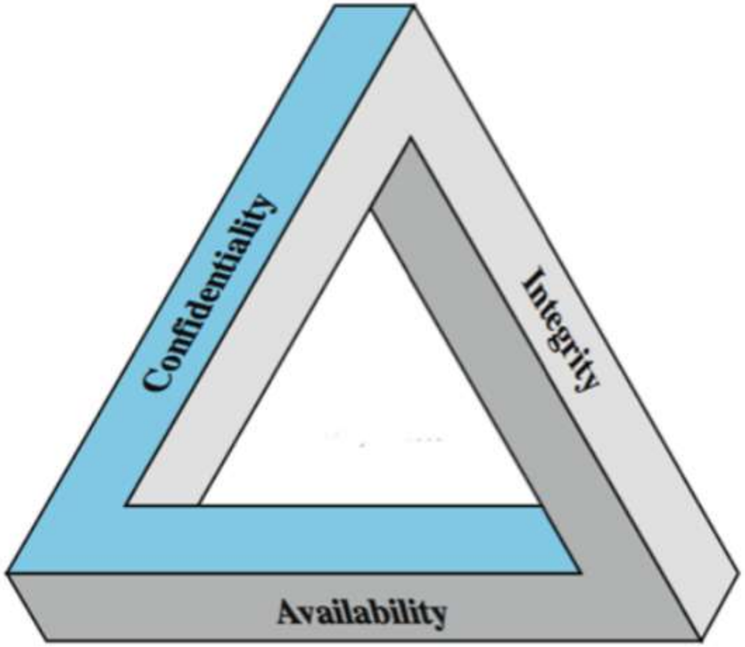
How to use spam filters to manipulate online elections

Hugo Jonker

Joint work with Sjouke Mauw (UL), Tom Schmitz (UL)

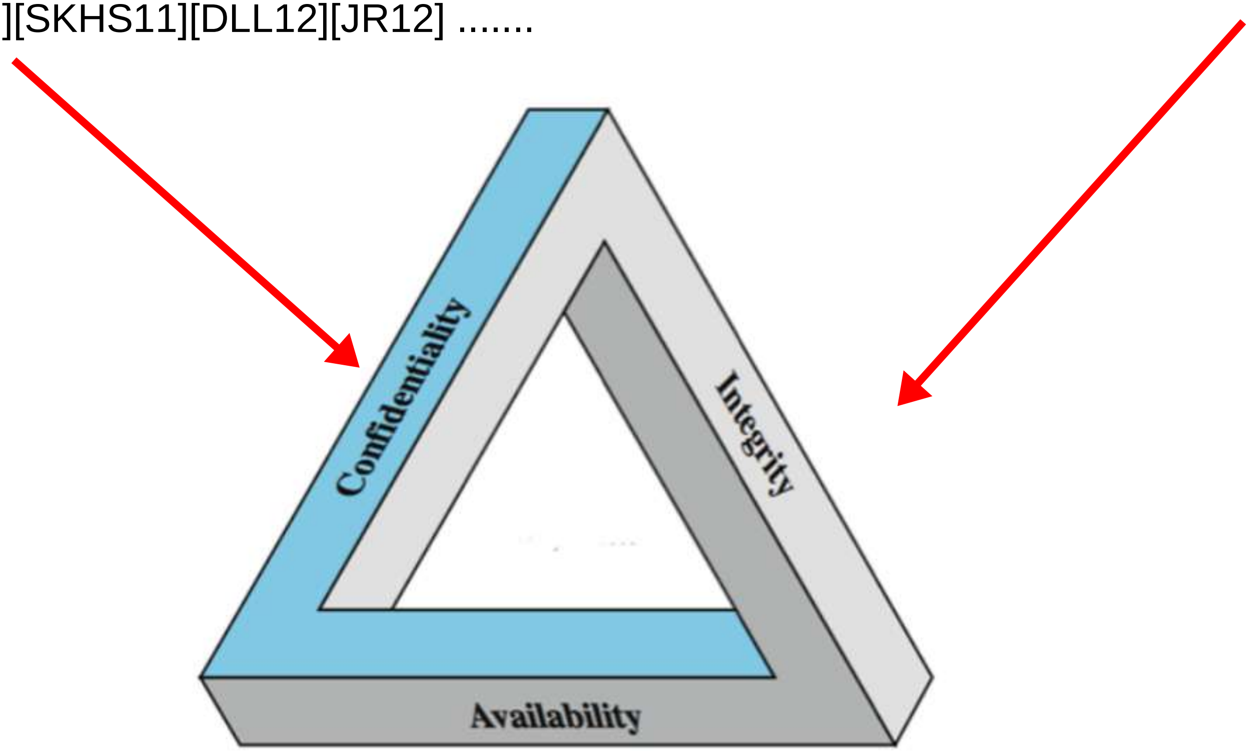
Open Universiteit
www.ou.nl

E-voting and the CIA triad



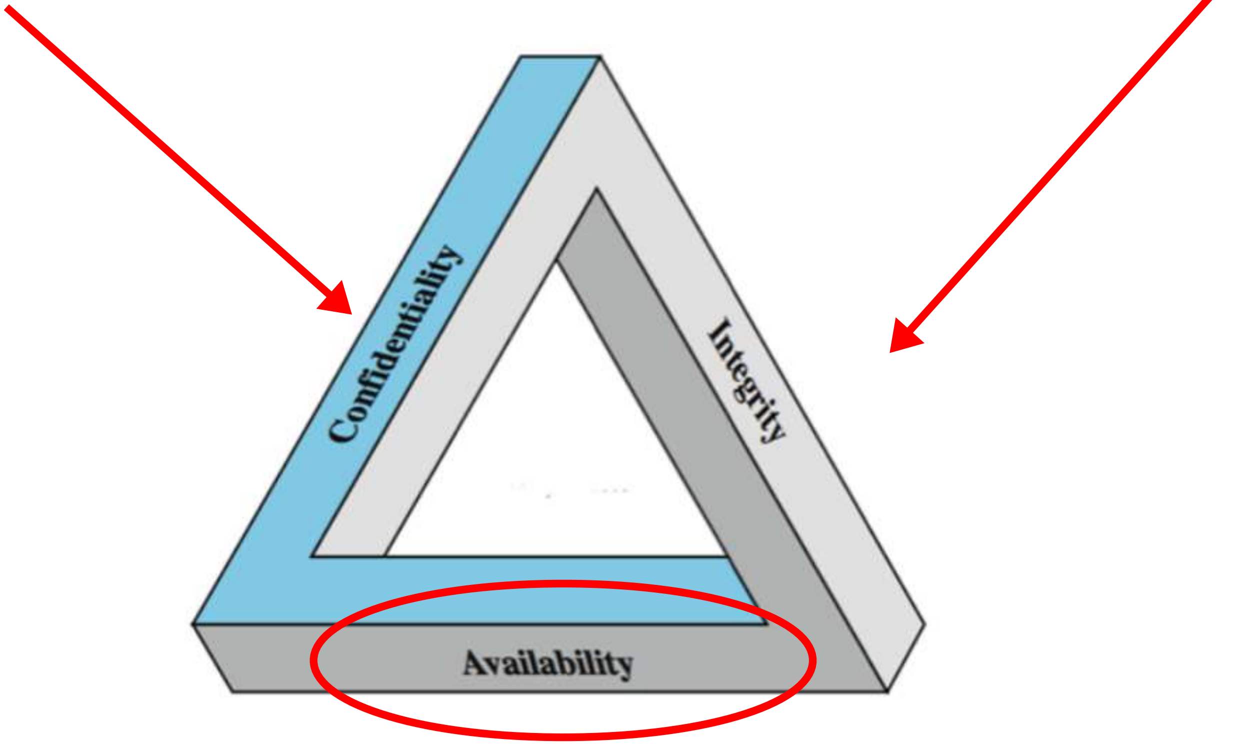
E-voting and the CIA triad

- [FOO92][BT94][NR94][SK95][Oka96][CSG97][HS00][LK00][Cha01][MBC01][Nef04][Cha04][Rya05][JCJ05][KR05][AN06][App06][AR06][FCS06][JdV06][MN06][Ben07][RS07][BMR07][MN07][HS07][Adi08][ECA08][PV08][JMP09][JRF09][ACvdG10][HRT10][KRMC10][KTV10][DLL11][JP11][SKHS11][DLL12][JR12] ....



E-voting and the CIA triad

- [FOO92][BT94][NR94][SK95][Oka96][CSG97][HS00][LK00][Cha01][MBC01][Nef04][Cha04][Rya05][JCJ05][KR05][AN06][App06][AR06][FCS06][JdV06][MN06][Ben07][RS07][BMR07][MN07][HS07][Adi08][ECA08][PV08][JMP09][JRF09][ACvdG10][HRT10][KRMC10][KTV10][DLL11][JP11][SKHS11][DLL12][JR12] ....



????

Literature on DoS attacks in e-voting

- Considered a serious threat
- Mostly ignored in security analysis
- Studied from a generic point of view
- Considered easily detectable
- Focuses on disruption of election process
  - Not on influencing outcome
- E-Vote-ID'17: using a DdoS prevention provider introduces new vulnerabilities

Attacker can manipulate election results if...

- DoS focused on selected voters
- Stealthy

Attacker can manipulate election results if...

- DoS focused on selected voters
- Stealthy

Reverse Bayesian Poisoning

Attacker can manipulate election results if...

- DoS focused on selected voters
- Stealthy

Reverse Bayesian Poisoning

Feasibility study on:

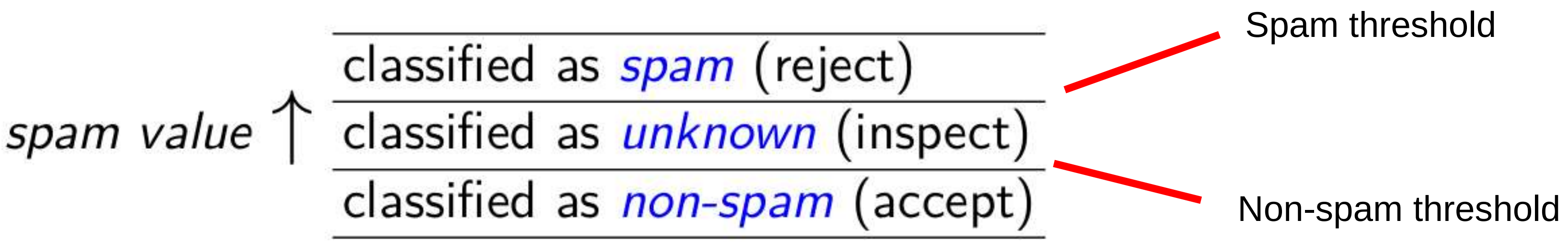
- Helios (voting system)
- BogoFilter (spam filter)

Spam

- > 50% of email traffic is spam
- Spam filtering is a necessity

Spam classification

- Spam filter calculates spam value of incoming message
- Based on word occurrence, URLs, sender, mime parts, ….



spam value ↑

classified as *spam* (reject)

classified as *unknown* (inspect)

classified as *non-spam* (accept)

Spam threshold

Non-spam threshold

Spam classification isn't perfect

- False accept:
  spam mail is marked as non-spam
- False reject:
  legitimate mail is discarded as spam

Bayesian spam filtering

Suppose incoming email contains “viagra”.

What is the probability that it is spam?

Bayes Theorem

IF it is known how often “viagra” occurs in

- spam: P(“viagra” | spam)
- and in non-spam: P(“viagra” | non-spam)

THEN we can compute this probability

Bayesian spam filtering

$$P(spam \mid \text{``}viagra'') =$$

$$\frac{P(\text{``}viagra'' \mid spam) \cdot P(spam)}{P(\text{``}viagra'' \mid spam) \cdot P(spam) + P(\text{``}viagra'' \mid \neg spam) \cdot P(\neg spam)}$$

Attacking Bayesian spam filters

- False accept
  spam mail considered legitimate

  Bayesian poisoning
  Add sufficiently many non-spam words to spam message

Attacking Bayesian spam filters

- False accept
  spam mail considered legitimate

  Bayesian poisoning
  Add sufficiently many non-spam words to spam message

- False reject
  legitimate mail discarded as spam

  Reverse Bayesian poisoning
  Train the spam filter with spam mails that also contain words from the regular mails that you want to be rejected

How to remotely train spam filter

Just send spam to the user.

Filter will be trained if:

- User marks incoming spam as spam
- Spam filter's auto-update feature is used

Feasibility experiment

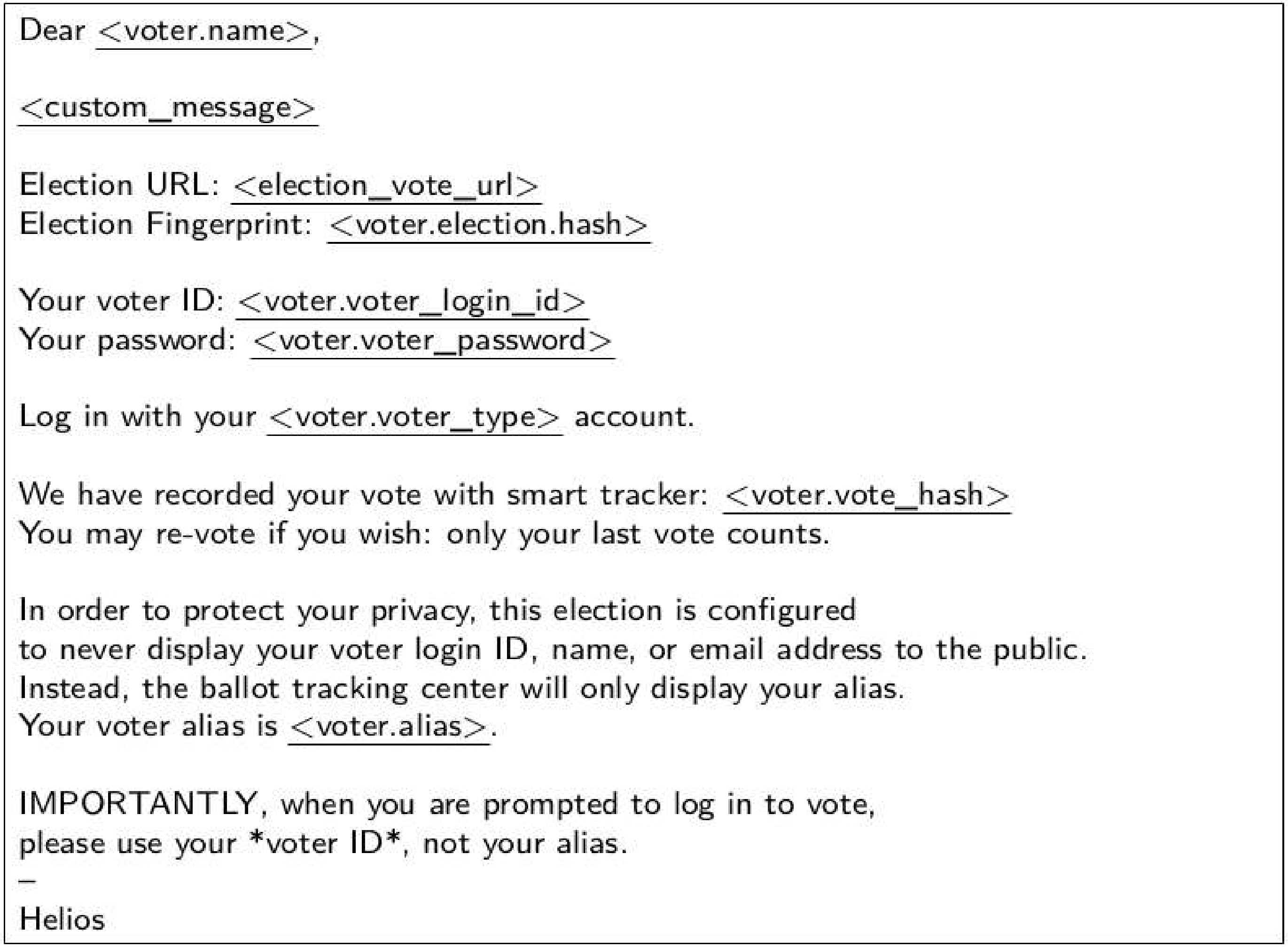
- Helios voting system
- Bogofilter Bayesian spam filter
- Trained with the Enron email corpus
- Fully controlled, isolated environment
  - No human interaction

Helios voting system

- online voting system
- Offers decent security / privacy guarantees
- Sends credentials via email

- We attack the email sending
  - Not part of the security guarantees
- If user doesn't vote, no action by Helios
- Cannot vote without these credentials

Template of a Helios invitation



Dear <voter.name>,

<custom_message>

Election URL: <election_vote_url>
Election Fingerprint: <voter.election.hash>

Your voter ID: <voter.voter_login_id>
Your password: <voter.voter_password>

Log in with your <voter.voter_type> account.

We have recorded your vote with smart tracker: <voter.vote_hash>
You may re-vote if you wish: only your last vote counts.

In order to protect your privacy, this election is configured
to never display your voter login ID, name, or email address to the public.
Instead, the ballot tracking center will only display your alias.
Your voter alias is <voter.alias>.

IMPORTANTLY, when you are prompted to log in to vote,
please use your *voter ID*, not your alias.
–
Helios

Example attack mail

From: Luxury@experience.com

Subject: Lower monthly payment passwords

Remuneration Election Subsidiary Link: payment Dear Usury – Reapportionment Helios Reply How Syndicate to Wholesale Vote Return ========== Computer Election roots URL: Coattail Your Challenger voter Believe ID: Decide Your Permit password: Advertisement Log Pamphlets in Broadcast with Downsize your…….

Example attack mail

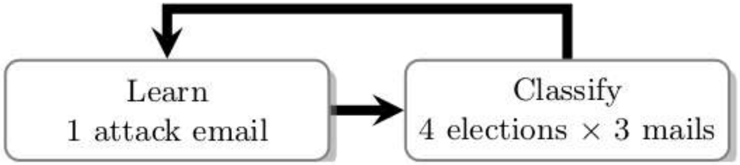
From: Luxury@experience.com

Subject: Lower monthly payment passwords

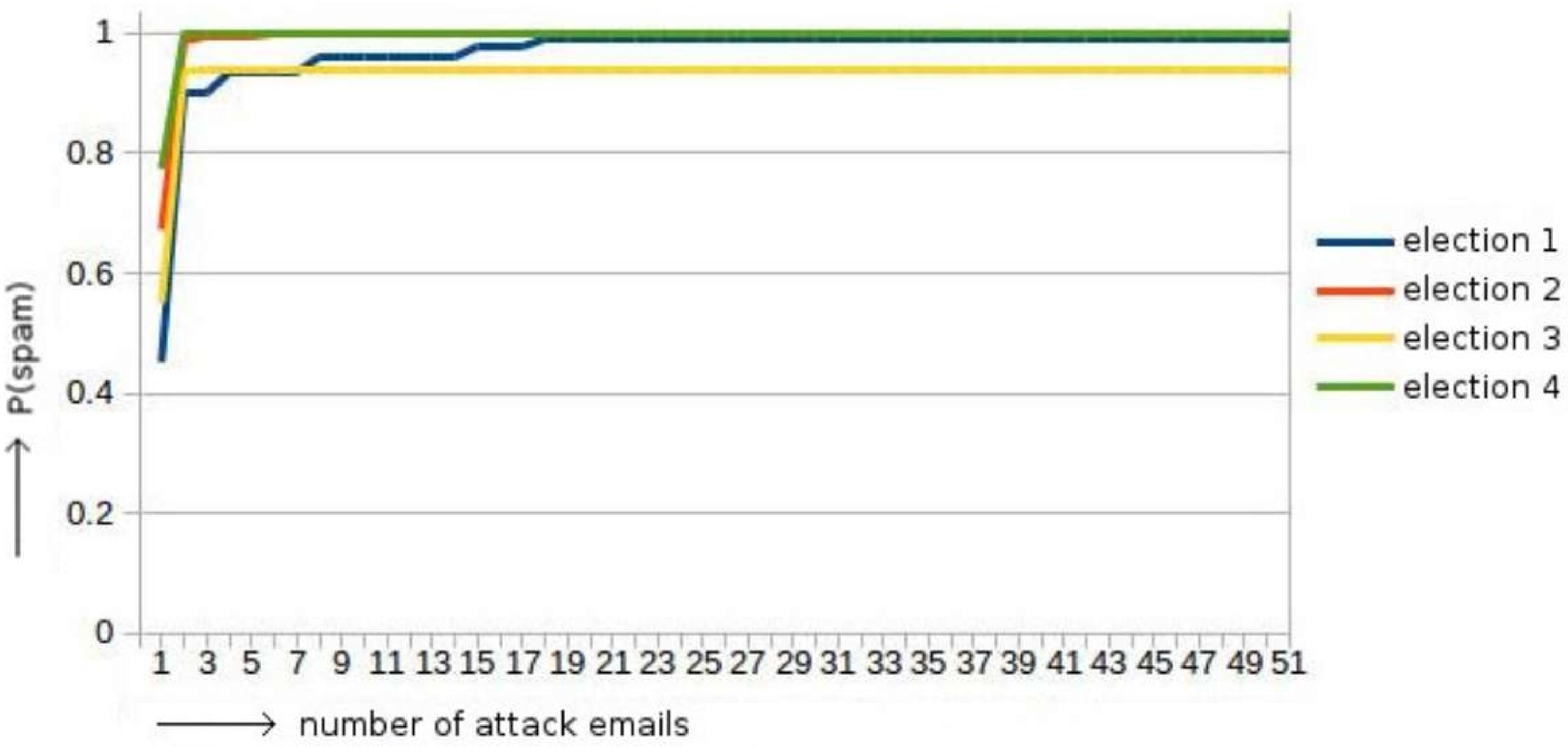
Remuneration **Election** Subsidiary **Link**: payment **Dear** Usury **–** Reapportionment **Helios** Reply **How** Syndicate **to** Wholesale **Vote** Return ========== Computer **Election** roots **URL:** Coattail **Your** Challenger **voter** Believe **ID:** Decide **Your** Permit **password:** Advertisement **Log** Pamphlets **in** Broadcast **with** Downsize **your**…….

Experiment

1. Set up 4 elections with Helios
2. Select 3 administrative emails per election
3. Train Bogofilter using Enron corpus
4. Train with 40 generated attack mails
   - One by one
5. After each attack mail:
   Classify admin emails



Results


1
0.8
0.6
0.4
0.2
0
P(spam)
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51
number of attack emails
election 1
election 2
election 3
election 4

Underlying assumptions

- Attacker knows which victims to attack
- Victims ignore lack of invitation for voting
- Active marking of spam or auto-update used
- Election officials don't train voters well enough to overcome this
- Elections are not repeated

Observations

- Just a few attack emails suffice to drastically increase spam value of genuine emails
- The attack is stealthy
  - Can be executed slowly
  - Evidence may even have been deleted if users retrain their spam filter

Future work

- Attacks can be optimised
  we showed feasibility
- Attack can be used against a group with shared spam filter
  - Requires field study with real subject
    Ethical considerations should prevent this
- Including emails and DoS attacks in formalisation and verification of voting protocols

Possible mitigations

- User-side:
  - Whitelisting election email address, calendar reminders for elections
- Central:
  - Multi-channel communication, request notification of receipt

Thank you for your attention!