Clustering Spam Campaigns with Fuzzy Hashing

Jianxing Chen, Romain Fontugne, Akira Kato, Kensuke Fukuda

Saarland University, National Institute of Informatics, JFLI, Keio University

Problems

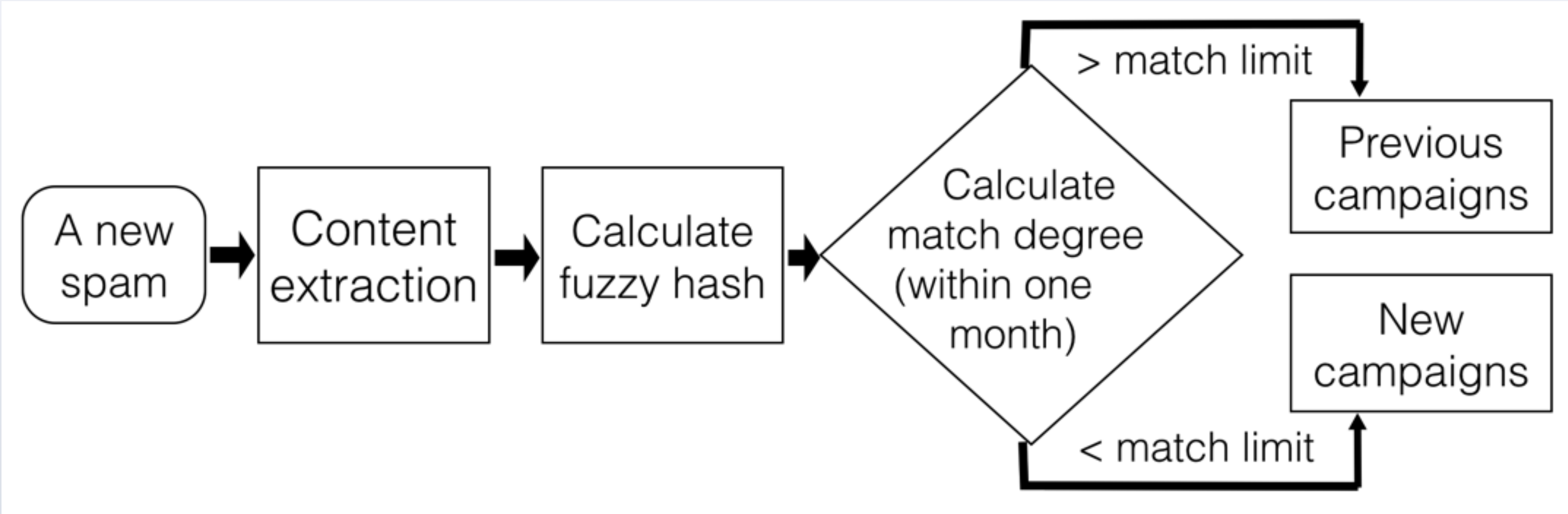
- **Goal: Identify spamming infrastructures**
- **Spammers send numerous emails in a stealthy manner using botnets**
- **Difficulties to identify spamming botnets**
  - Each bot sends a small number of spam emails
  - Bots are spread worldwide
  - Spam campaigns last for months

Proposed Approach

- **Infer botnets from spam campaigns**
- **Identification of spam campaigns?**
  - Find spams with common tokens?
    ⇒ Easily evaded with obfuscated techniques (e.g. URL shortening)
  - Find spams serving a common purpose
    ⇒ **Cluster spams content with fuzzy hash!**

Methodology

1. **Feature extraction (tokens, email body, title, …)**
2. **Campaign clustering**
   using fuzzy hashing
3. **Botnet inference**
   using SMTP servers path


A new spam → Content extraction → Calculate fuzzy hash → Calculate match degree (within one month)

> match limit → Previous campaigns

< match limit → New campaigns

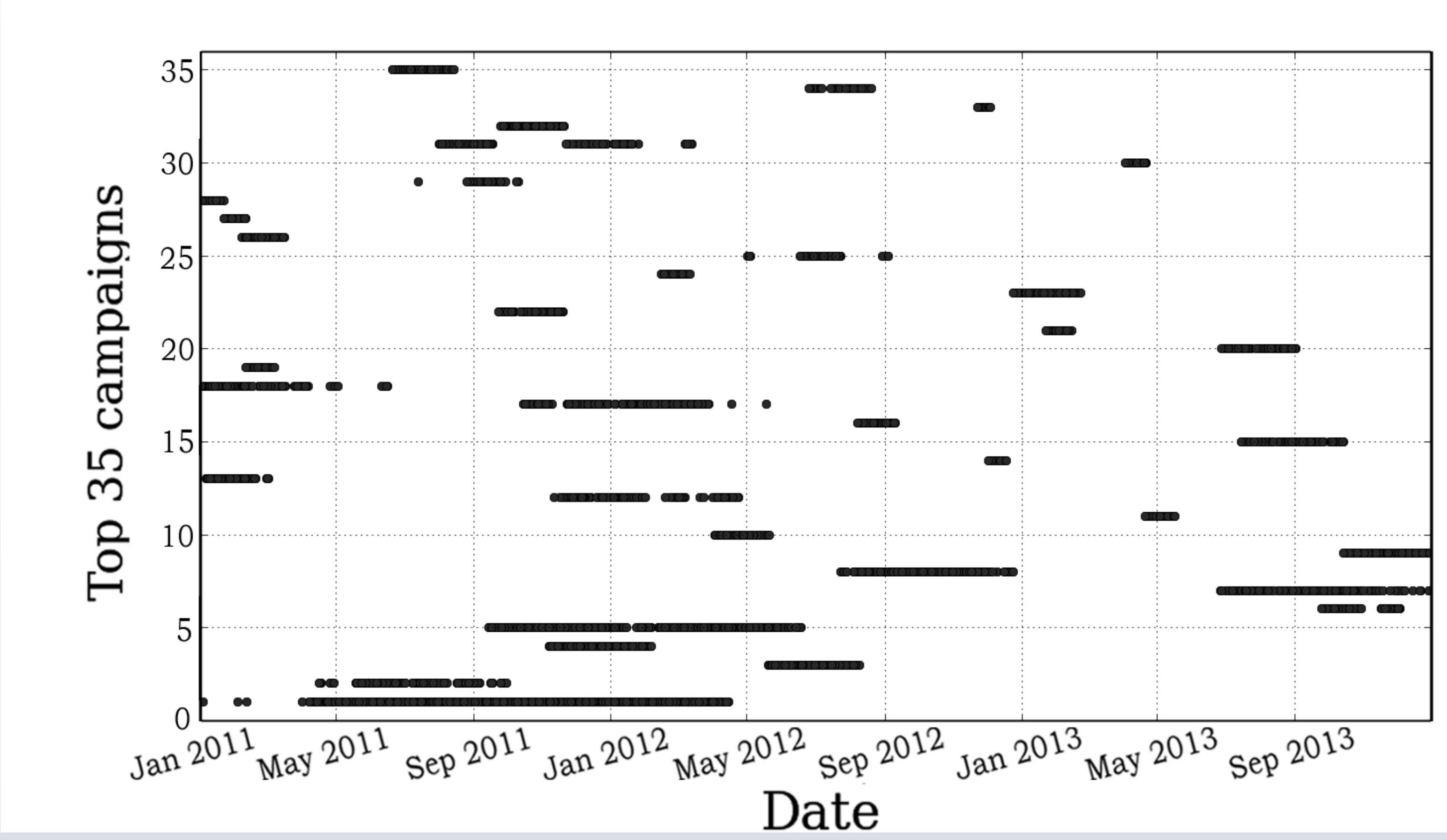
Fuzzy hashing:

- Based on 2 hash functions
- Compute hashes that are comparable with weighted edit distance (match degree)

Results

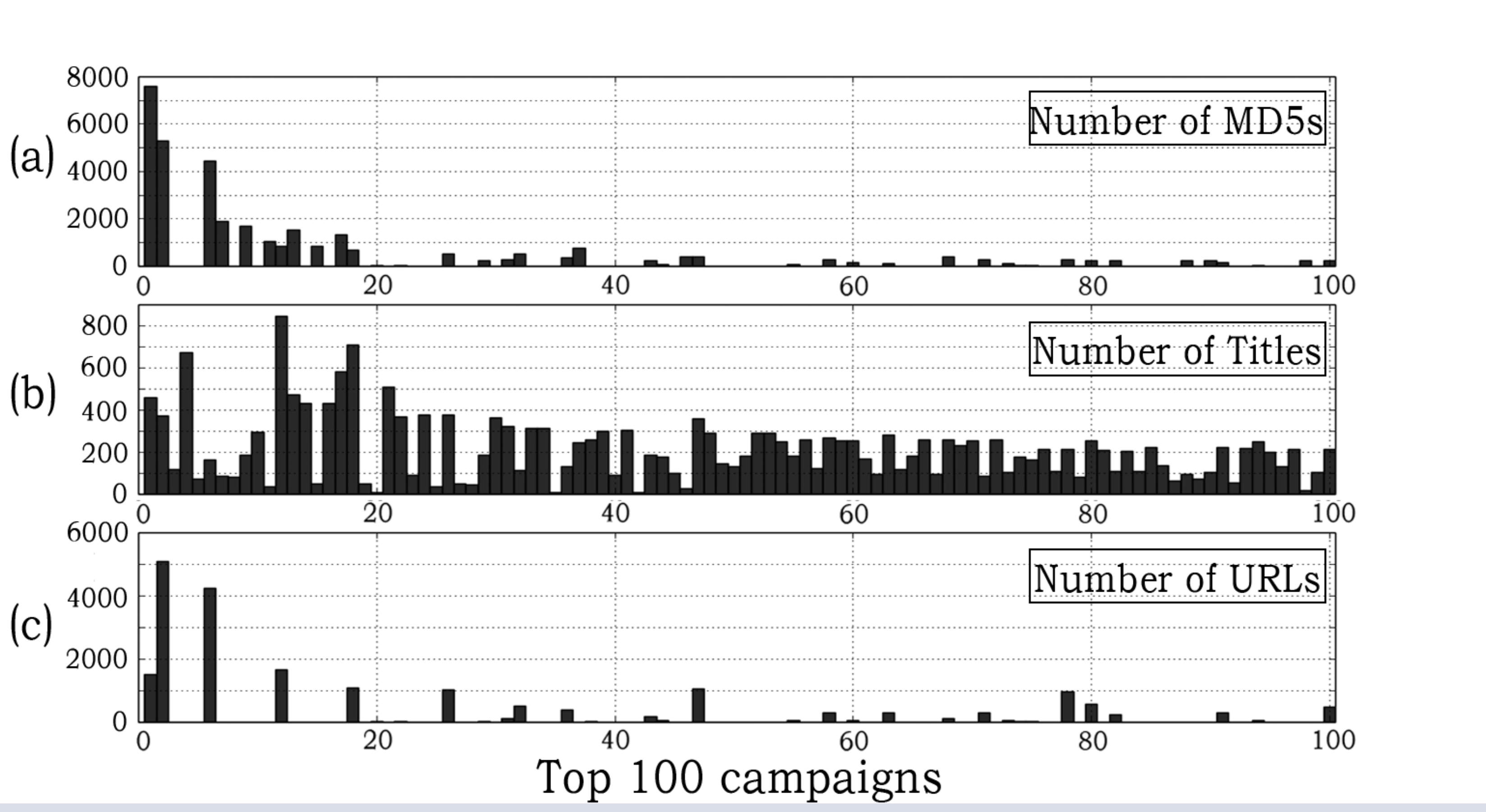
- **Dataset:** 540k spam emails from a few accounts.

Time evolution of top 35 campaigns:



⇒ campaigns usually last a few months

Content characteristics of top 100 campaigns:



⇒ MD5: 2 types of campaigns, URL: good token

Reference

J. Chen et al. Clustering Spam Campaigns with Fuzzy Hashing, 10th Asian Internet Engineering Conference (AINTEC 2014), Nov.26-28, 2014.

Romain FONTUGNE,
romain@nii.ac.jp