SEVEN MONTHS' WORTH OF MISTAKES

A Longitudinal Study of Typosquatting Abuse

NDSS, February 9th 2015

Pieter Agten Wouter Joosen Frank Piessens Nick Nikiforakis*

iMinds-DistriNet, KU Leuven, Belgium

* **Stony Brook University**, New York, USA

BACKGROUND

WHAT IS TYPOSQUATTING?

``Typosquatting is the act of purposefully registering a domain name that is a mistype of another domain name.''

WHAT IS TYPOSQUATTING?

``Typosquatting is the act of purposefully registering a domain name that is a mistype of another domain name.''

For example `youtuve.com` instead of `youtube.com`

WHAT IS TYPOSQUATTING?

``Typosquatting is the act of purposefully registering a domain name that is a mistype of another domain name.''

For example `youtuve.com` instead of `youtube.com`

Why?

- Anticipate typos in browser's address bar
- In order to steal hits from legitimate websites

OUR STUDY: BACKGROUND

Typosquatting has been known and studied for 15+ years

OUR STUDY: BACKGROUND

Typosquatting has been known and studied for 15+ years, but no *content-based* longitudinal studies have been published.

OUR STUDY: BACKGROUND

Typosquatting has been known and studied for 15+ years, but no *content-based* longitudinal studies have been published.

We studied the typosquatting domains of the 500 most popular websites **over a period of 7 months**.

OUR STUDY: BACKGROUND

Typosquatting has been known and studied for 15+ years, but no *content-based* longitudinal studies have been published.

We studied the typosquatting domains of the 500 most popular websites **over a period of 7 months**.

Longitudinal in addition to static observations:

- Are typosquatting domains changing hands?
- Does their content change over time?

OUR STUDY: APPROACH

1. Take Alexa top 500

OUR STUDY: APPROACH

1. Take Alexa top 500

2. Generate all potential typosquatting domains

OUR STUDY: APPROACH

1. Take Alexa top 500

2. Generate all potential typosquatting domains

   Based on models of [Wang-2006]

   - Missing dot `wwwexample.com`
   - Character-omission `www.exmple.com`
   - Character-permutation `www.examlpe.com`
   - Character-substitution `www.ezample.com`
   - Character-duplication `www.exaample.com`

OUR STUDY: APPROACH

1. Take Alexa top 500

2. Generate all potential typosquatting domains

   28,179 potential
   17,172 active

   Based on models of [Wang-2006]

   - Missing dot — `wwwexample.com`
   - Character-omission — `www.exmple.com`
   - Character-permutation — `www.examlpe.com`
   - Character-substitution — `www.ezample.com`
   - Character-duplication — `www.exaample.com`

OUR STUDY: APPROACH

1. Take Alexa top 500
2. Generate all potential typosquatting domains
3. From April 1st, 2013 until October 31st, 2013:
   - Visit each domain daily using PhantomJS
   - Download WHOIS information weekly

28,179 potential
17,172 active

OUR STUDY: APPROACH

1. Take Alexa top 500

2. Generate all potential typosquatting domains

3. From April 1st, 2013 until October 31st, 2013:
   - Visit each domain daily using PhantomJS
   - Download WHOIS information weekly

28,179 potential
17,172 active

3.4 million visits
900 GB

OUR STUDY: APPROACH

1. Take Alexa top 500

2. Generate all potential typosquatting domains

28,179 potential
17,172 active

3. From April 1st, 2013 until October 31st, 2013:
   - Visit each domain daily using PhantomJS
   - Download WHOIS information weekly

3.4 million visits
900 GB

4. Cluster and categorize collected pages
   - Initial automatic clustering based on visual similarity
   - Second-stage manual validation and content-based categorization

OUR STUDY: APPROACH

1. Take Alexa top 500

2. Generate all potential typosquatting domains

28,179 potential
17,172 active

3. From April 1st, 2013 until October 31st, 2013:
   - Visit each domain daily using PhantomJS
   - Download WHOIS information weekly

3.4 million visits
900 GB

4. Cluster and categorize collected pages
   - Initial automatic clustering based on visual similarity
   - Second-stage manual validation and content-based categorization

8,102 clusters

OUR STUDY: APPROACH

1. Take Alexa top 500

2. Generate all potential typosquatting domains

28,179 potential
17,172 active

3. From April 1st, 2013 until October 31st, 2013:
   - Visit each domain daily using PhantomJS
   - Download WHOIS information weekly

3.4 million visits
900 GB

4. Cluster and categorize collected pages
   - Initial automatic clustering based on visual similarity
   - Second-stage manual validation and content-based categorization

8,102 clusters

5. Analyze categorized clusters

RESULTS

MALICIOUS VS DEFENSIVE REGISTRATIONS

95% of the Alexa Top 500 have at least 1 malicious typo domain

Most malicious registrations: `adultfriendfinder.com` (132)

MALICIOUS VS DEFENSIVE REGISTRATIONS

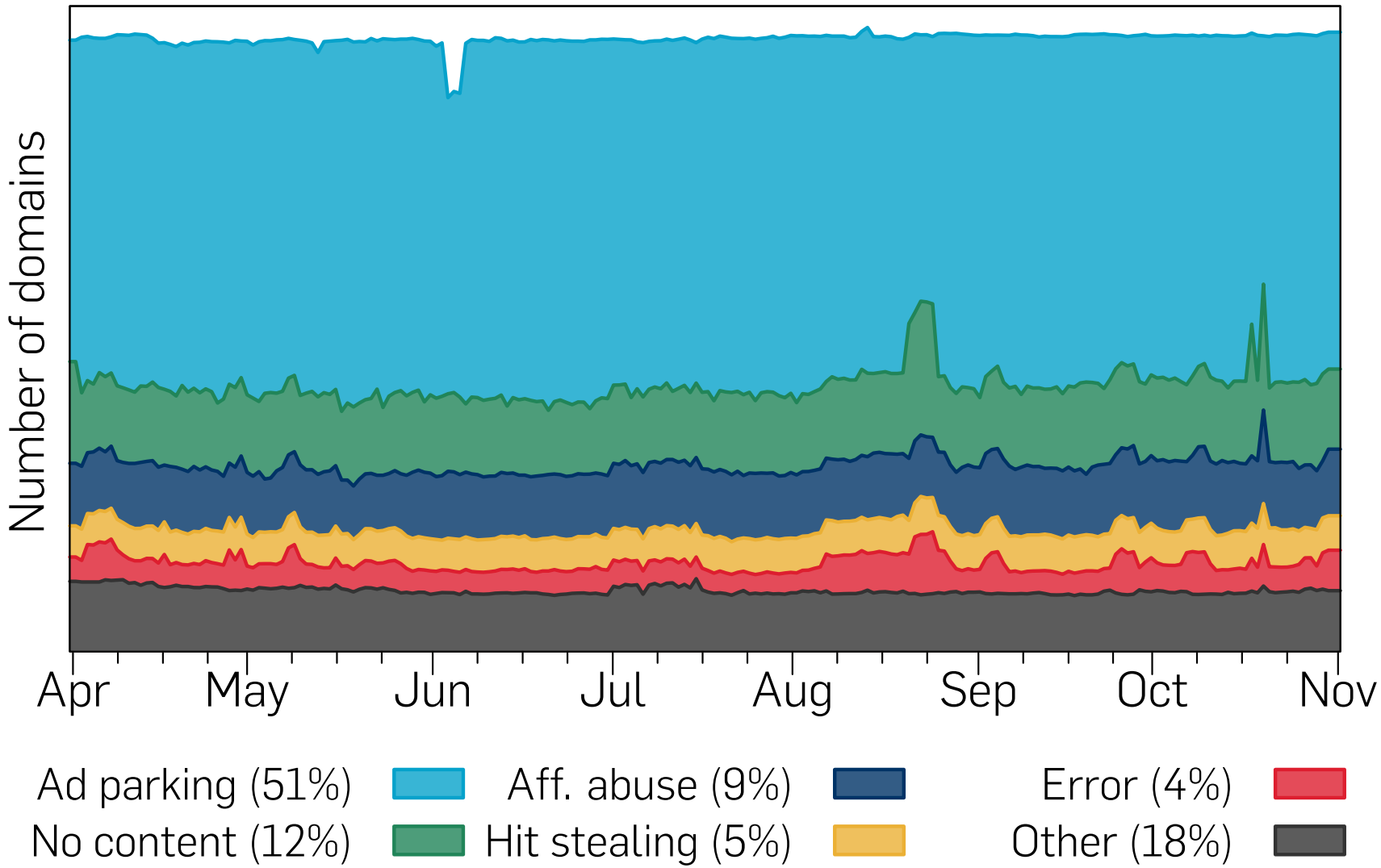
95% of the Alexa Top 500 have at least 1 malicious typo domain

Most malicious registrations: `adultfriendfinder.com` (132)

31% of the Alexa Top 500 have at least 1 defensive registration

Most defensive registrations: `huffingtonpost.com` (57)

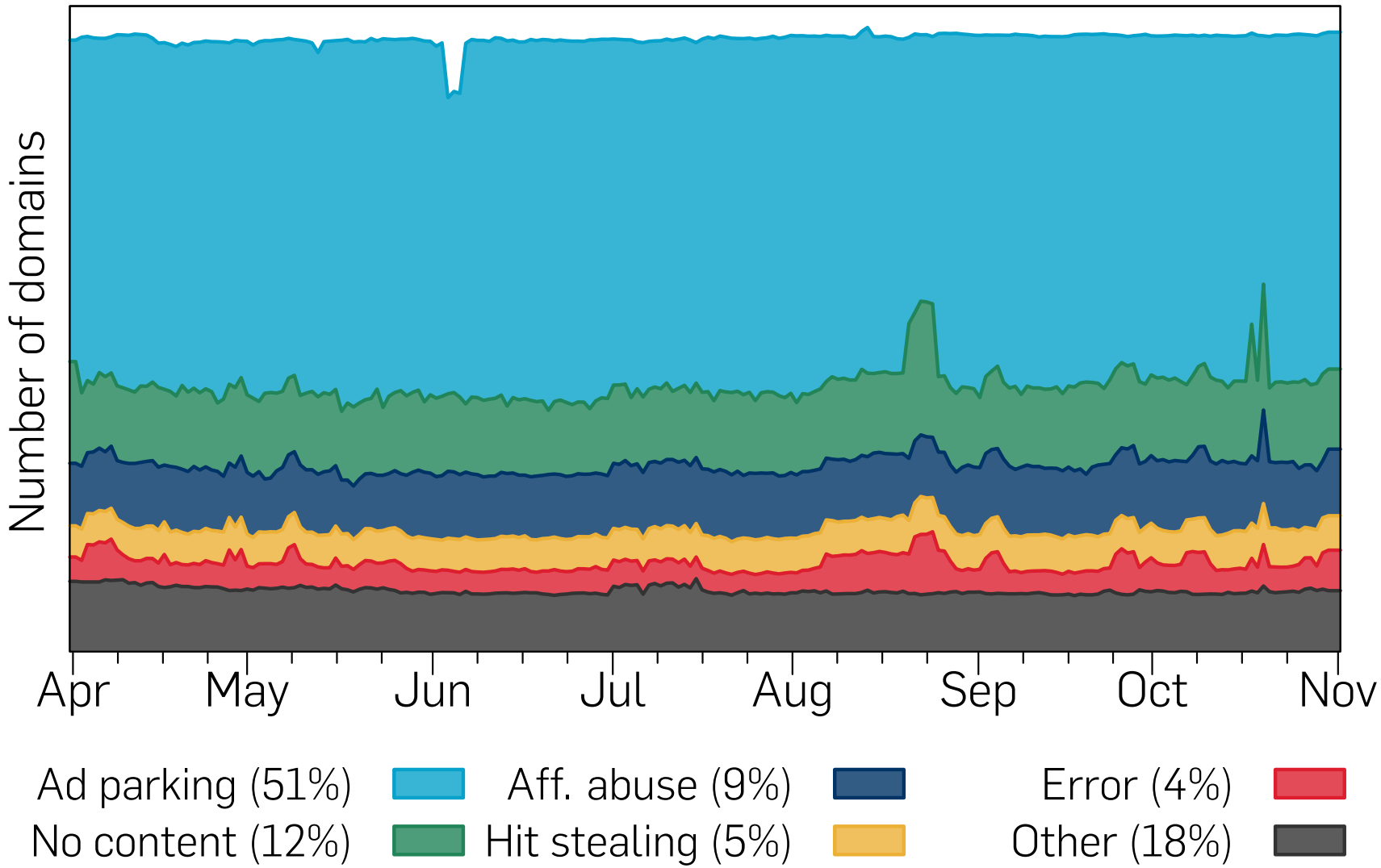
CATEGORY DISTRIBUTION OVER TIME



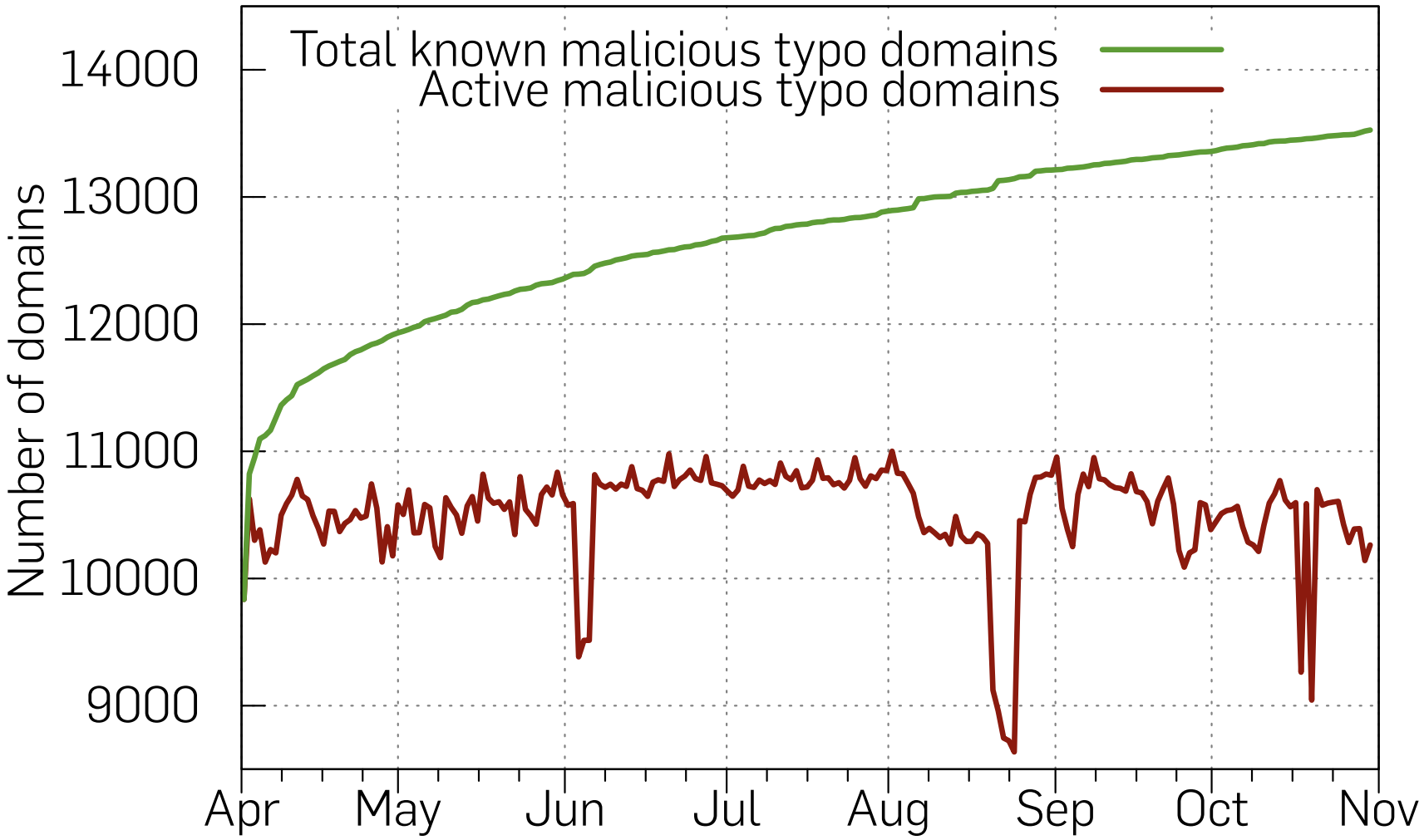
CATEGORY DISTRIBUTION OVER TIME

Ad parking (51%)	Aff. abuse (9%)	Error (4%)
No content (12%)	Hit stealing (5%)	Other (18%)

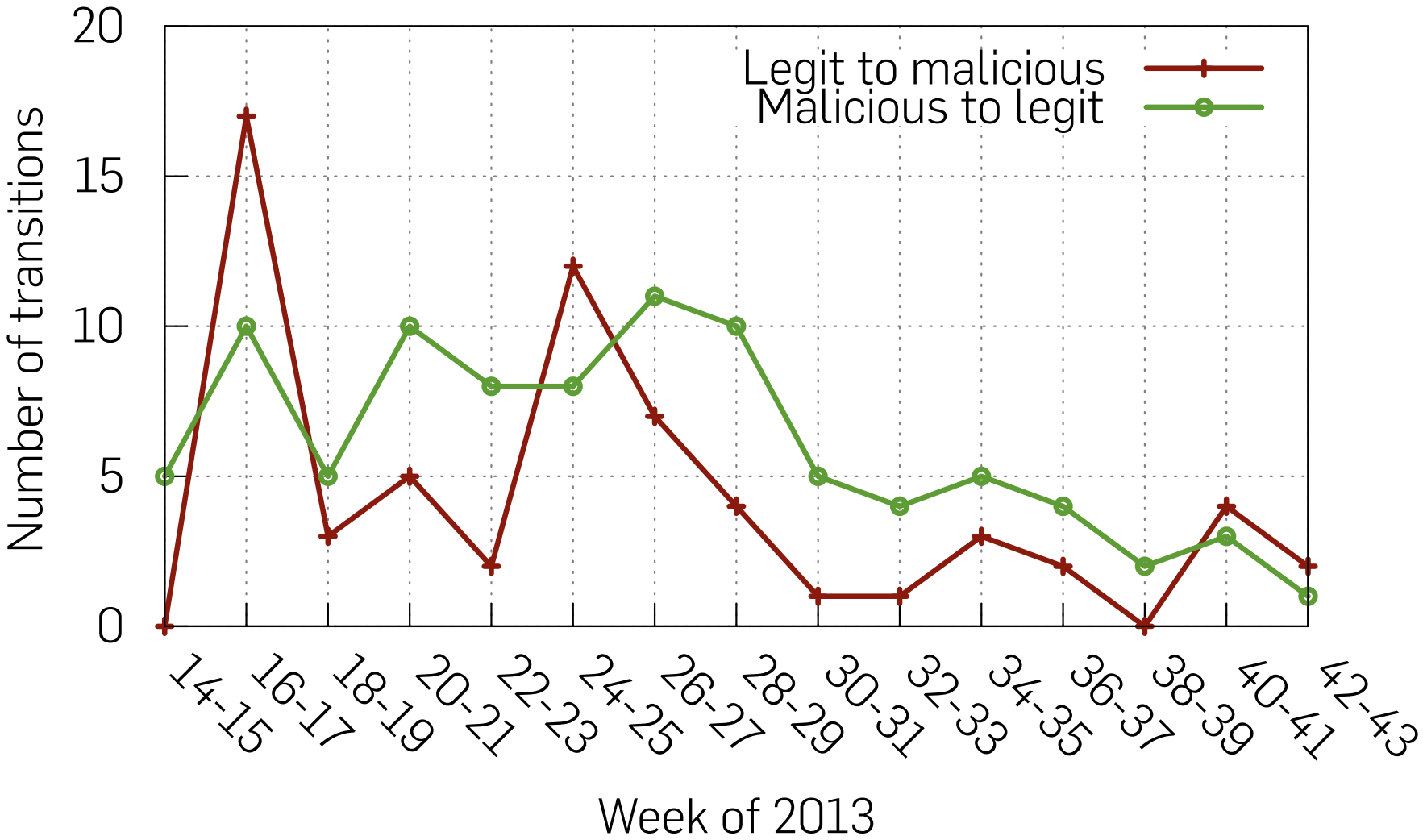
CATEGORY DISTRIBUTION OVER TIME



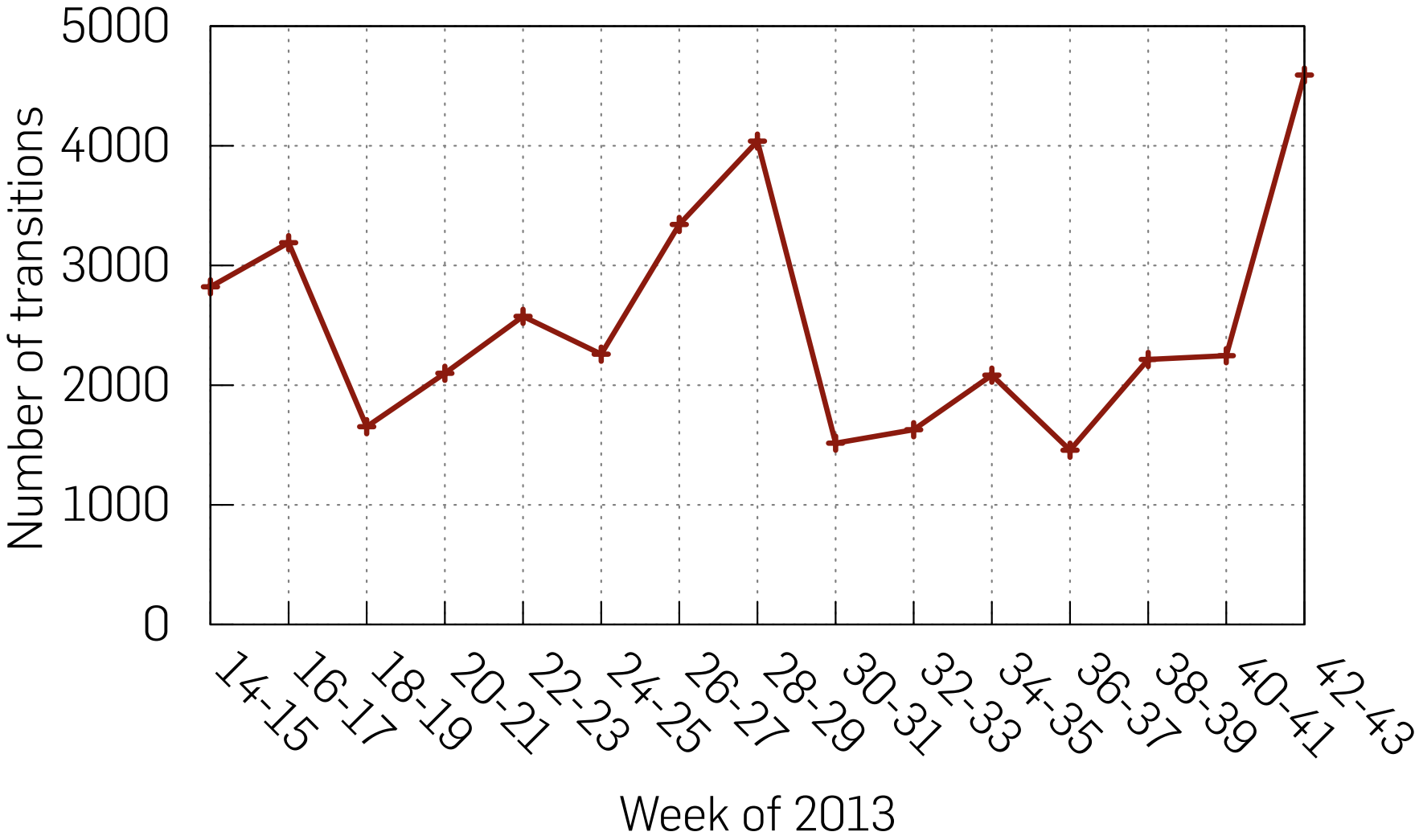
NUMBER OF TYPOSQUATTING DOMAINS OVER TIME



NUMBER OF LEGIT-MALICIOUS TRANSITIONS OVER TIME



NUMBER OF CATEGORY TRANSITIONS OVER TIME



LARGEST TYPOSQUATTING PAGE HOSTERS

Table: Number of malicious domains per autonomous system

Network	Registered owner	Malicious domains
`208.73.210.0 /23`	Oversee.net	2,405 (18%)
`199.59.243.96 /28`	Bodis	1,741 (13%)
`82.98.86.160 /27`	Sedo	1,388 (10%)
`69.43.161.128 /25`	Castle Access	1,216 (9%)
Total		**6,750 (50%)**

CONCLUSION

CONCLUSION

We performed the first content-based longitudinal typosquatting study, which shows that:

- Typosquatting is still actively practiced today
- Many malicious, but only few defensive registrations
- Typo domains are changing hands and changing content
- Only 4 hosters account for 50% of the typosquatting domains

Many other results can be found in the paper

Our entire dataset is available at
https://distrinet.cs.kuleuven.be/software/typos15/