nominet

and Oxford Brookes University

A Statistical Approach to Typosquatting Detection

Alessandro Linari
alessandro@nominet.org.uk

DNS Ops Workshop, 4-5 June 2008

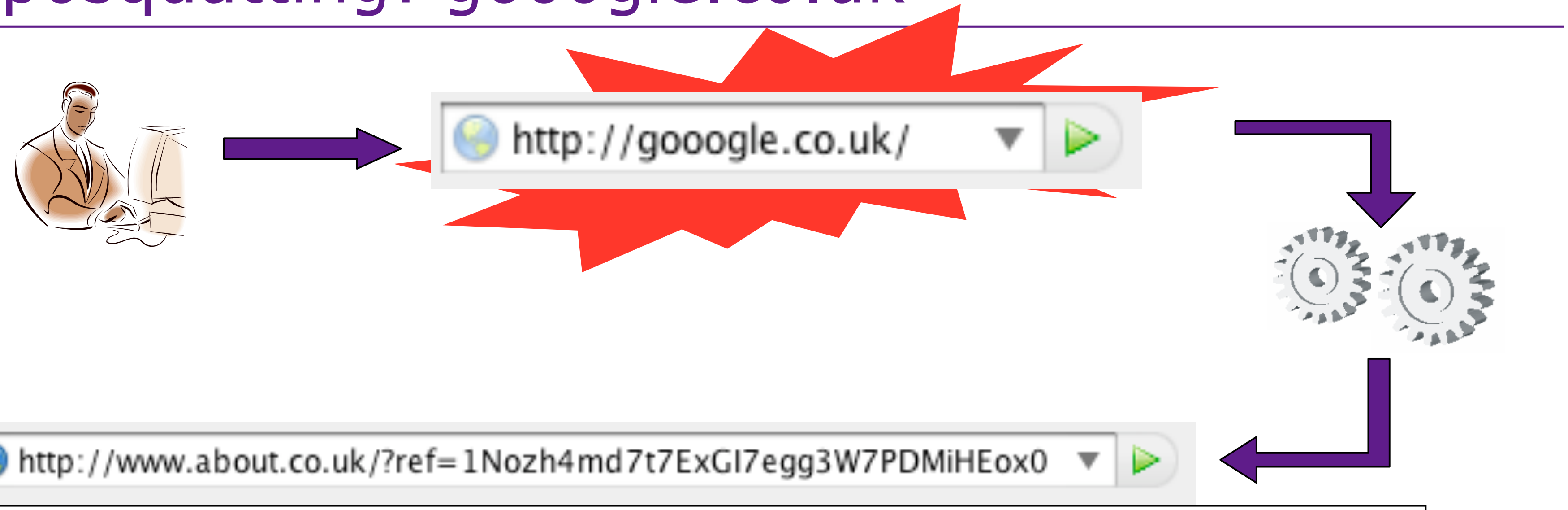
Introduction

Typosquatting is the practice of registering a domain name which contains a typographical error if compared to the name of a trademark or a famous domain

- Growing phenomenon over the Internet
  - Well-understood from a legal point of view
  - Lack of a technical characterisation
- First attempt for
  - Technical definition
  - Statistical characterisation

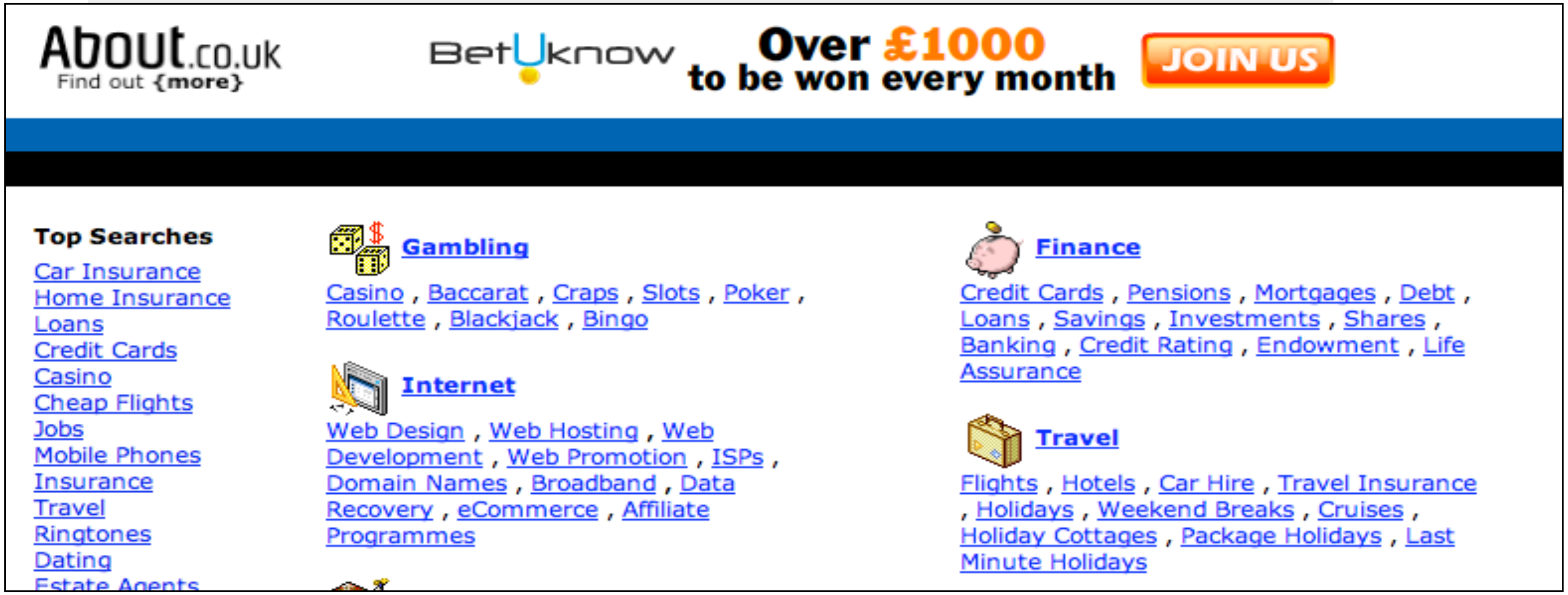
nominet

Typosquatting: gooogle.co.uk



http://gooogle.co.uk/

http://www.about.co.uk/?ref=1Nozh4md7t7ExGI7egg3W7PDMiHEox0



About.co.uk Find out {more}

BetUknow **Over £1000 to be won every month** JOIN US

Top Searches

Car Insurance
Home Insurance
Loans
Credit Cards
Casino
Cheap Flights
Jobs
Mobile Phones
Insurance
Travel
Ringtones
Dating
Estate Agents

Gambling

Casino , Baccarat , Craps , Slots , Poker , Roulette , Blackjack , Bingo

Internet

Web Design , Web Hosting , Web Development , Web Promotion , ISPs , Domain Names , Broadband , Data Recovery , eCommerce , Affiliate Programmes

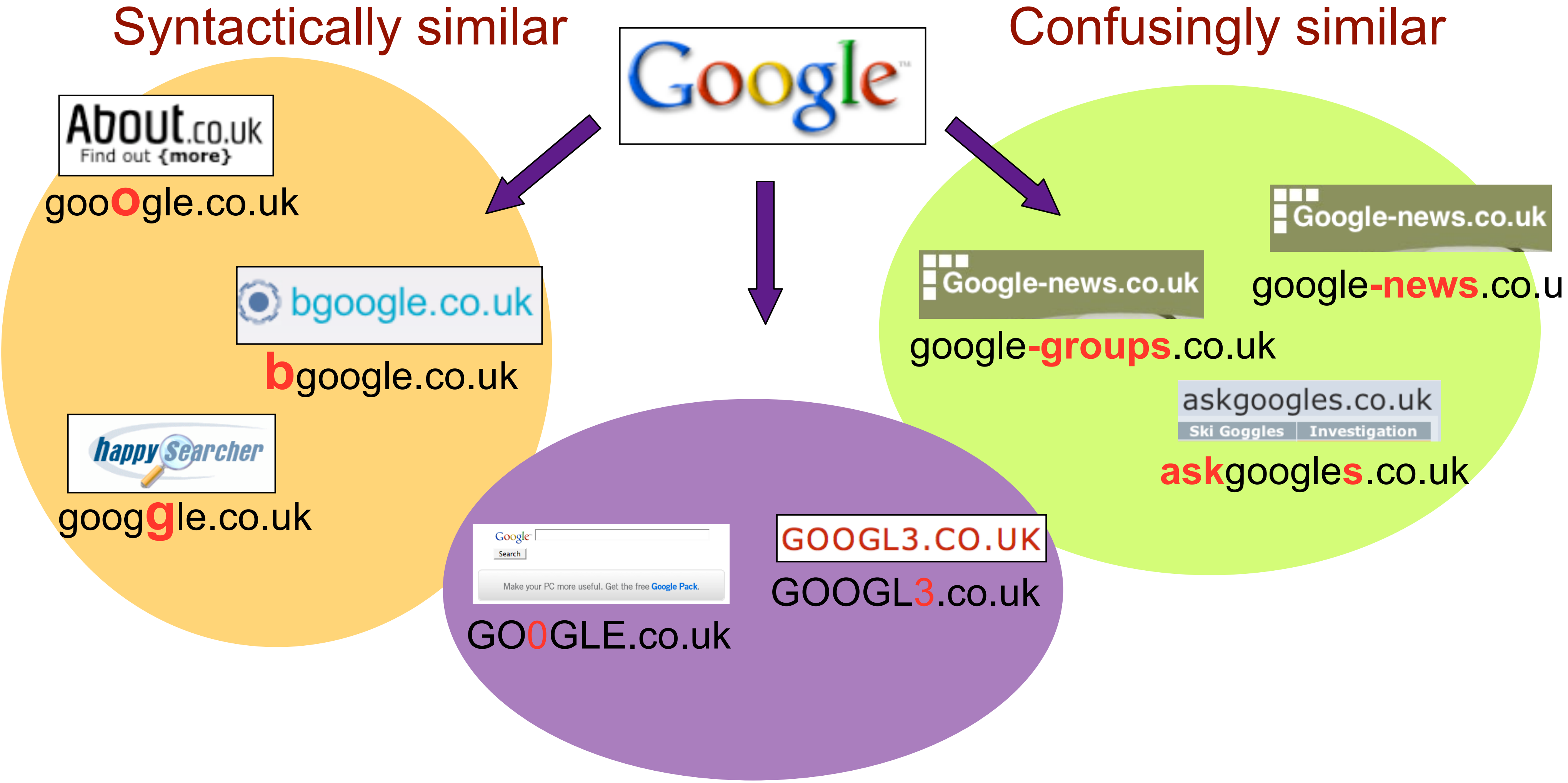
Finance

Credit Cards , Pensions , Mortgages , Debt , Loans , Savings , Investments , Shares , Banking , Credit Rating , Endowment , Life Assurance

Travel

Flights , Hotels , Car Hire , Travel Insurance , Holidays , Weekend Breaks , Cruises , Holiday Cottages , Package Holidays , Last Minute Holidays

Syntactic and Confusing Similarity



Syntactically similar

goooogle.co.uk

bgoogle.co.uk

googgle.co.uk

Confusingly similar

google-news.co.u

google-groups.co.uk

askgoogles.co.uk

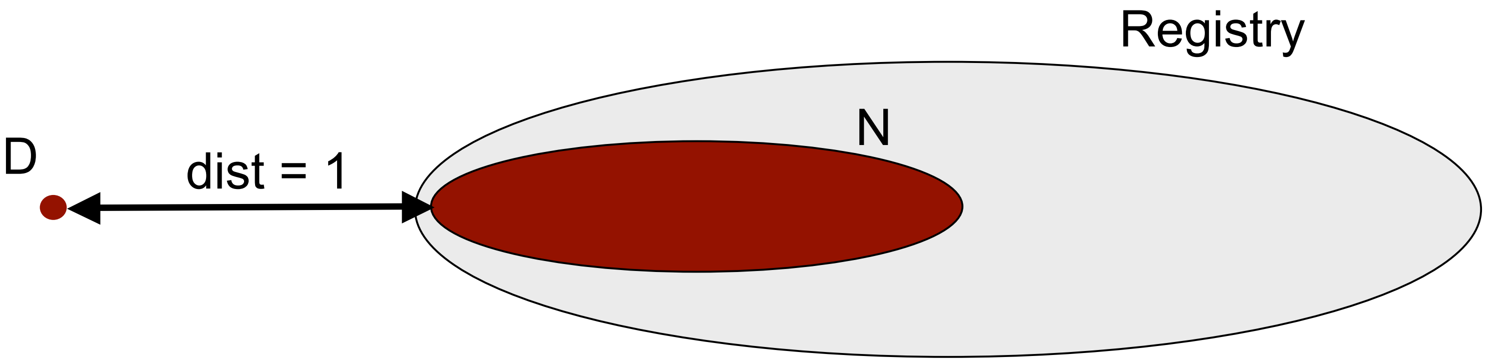
GOOGL3.co.uk

GO0GLE.co.uk

Visually similar

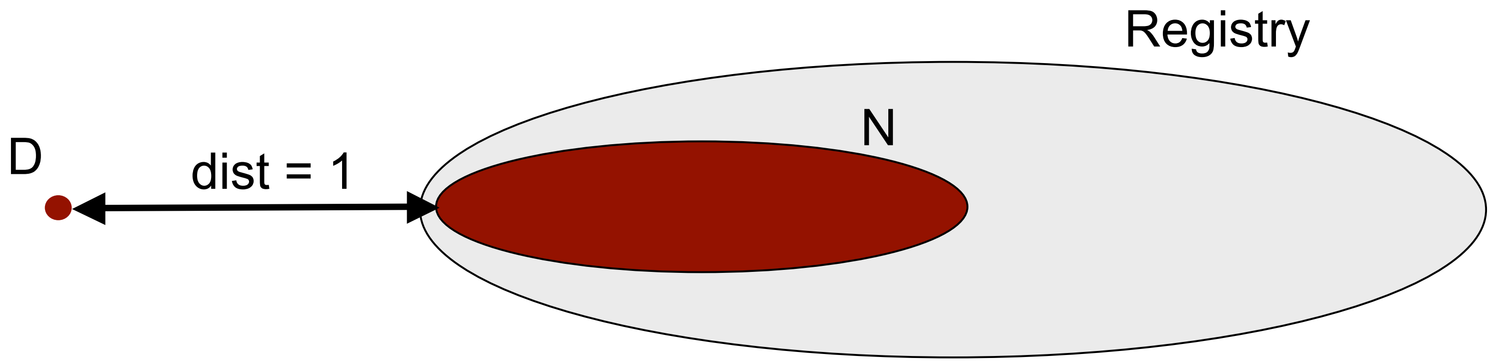
Syntactic Neighbourhood

Given a domain D, the syntactic neighbourhood of D set of all domains in the registry whose edit distance from D is equal to 1



Syntactic Neighbourhood

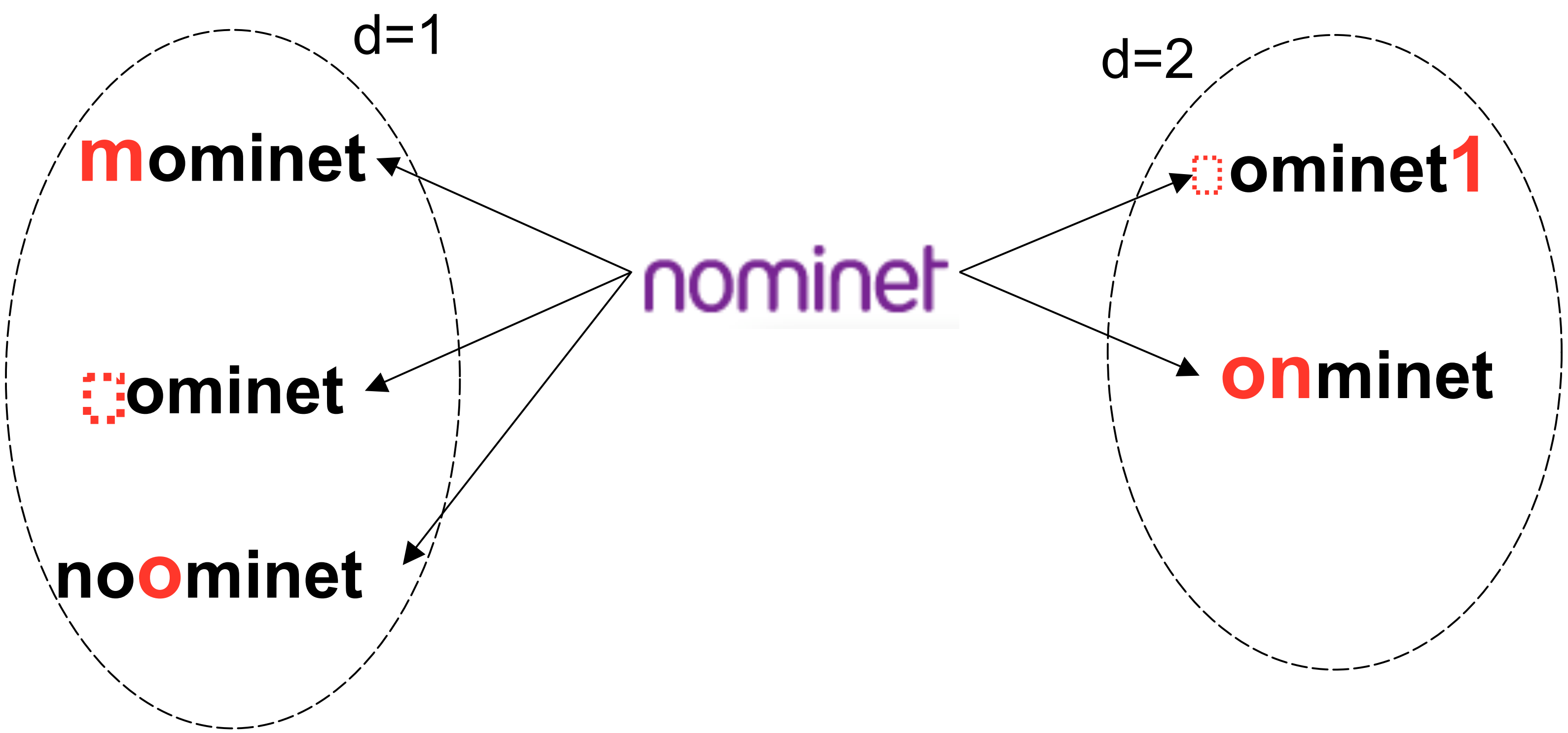
Given a domain D, the syntactic neighbourhood of D set of all domains in the registry whose edit distance from D is equal to 1



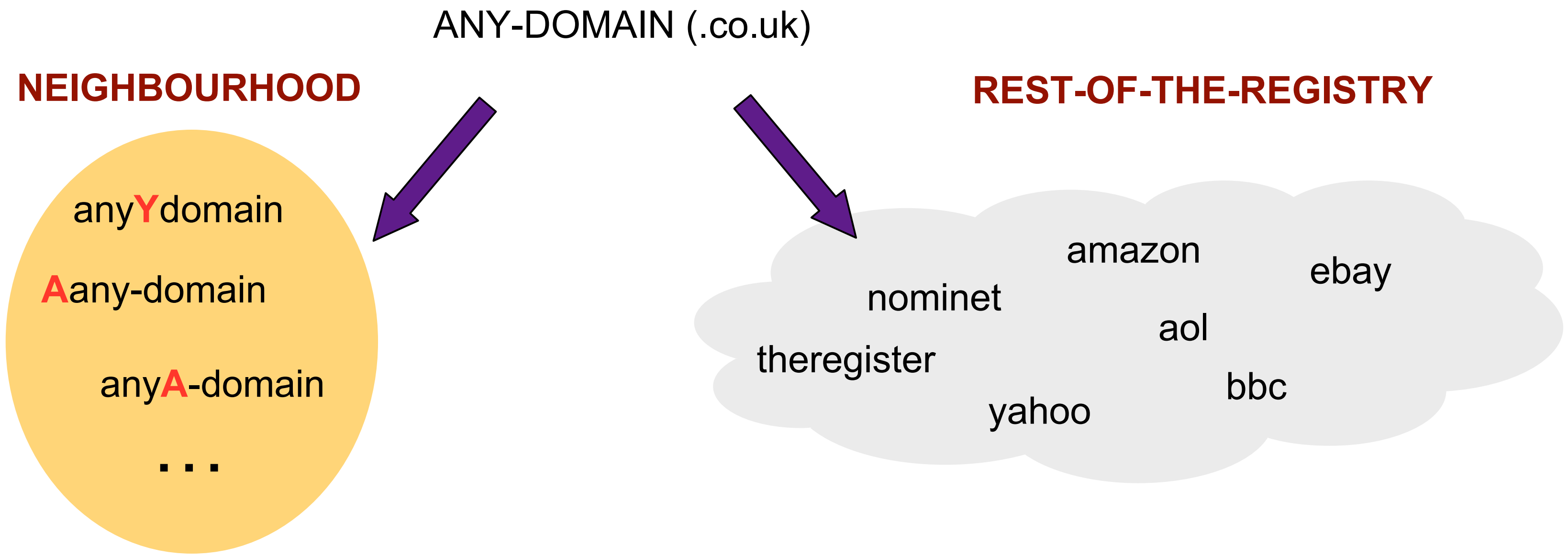
- Edit distance
  - Minimum number of operations needed to transform one string into the other
  - An operation is an insertion, deletion, or substitution of a single character

Syntactic Neighbourhood

Given a domain D, the syntactic neighbourhood of D set of all domains in the registry whose edit distance from D is equal to 1

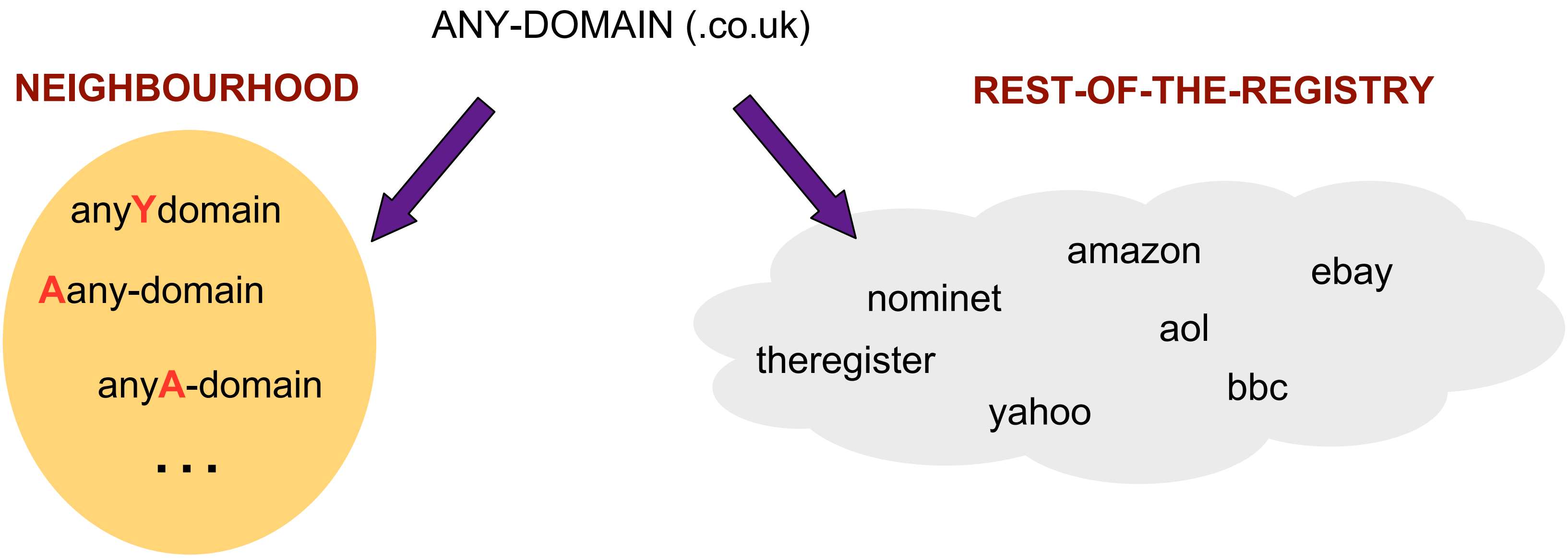


Outline



- Correlation between popularity of a domain name and size of its neighbourhood
- Presence of "typosquatters friendly" registrars in the neighbourhood of popular domains

Outline



- Correlation between popularity of a domain name and size of its neighbourhood
- Presence of "typosquatters friendly" registrars in the neighbourhood of popular domains

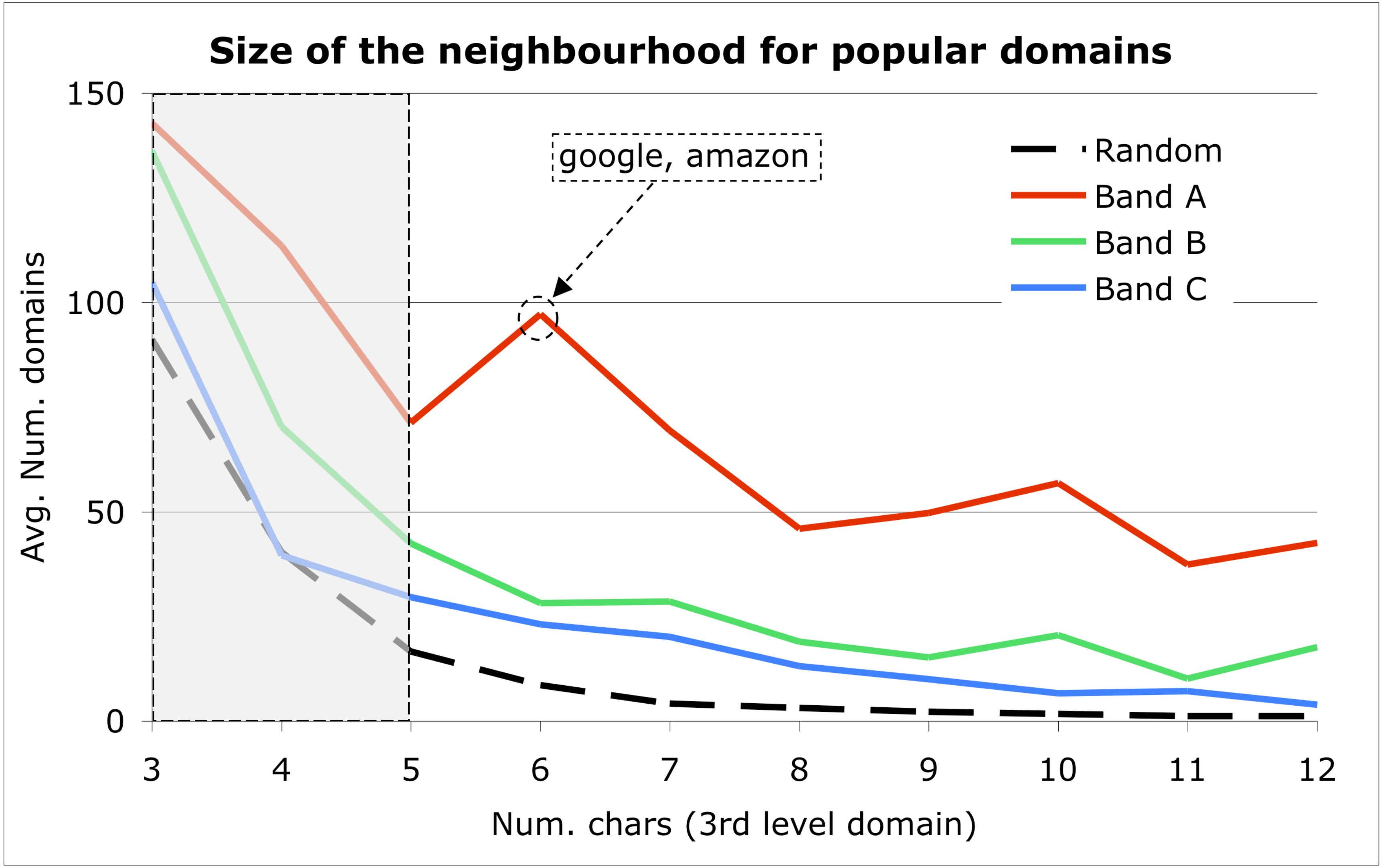
Experimental Setting

- Choose a domain name X
- Compute the distance between X and all domains in the registry
- Compute the size of X's neighbourhood
- Compute the average size of a neighbourhood for domains of each length
    - E.g., bbc.co.uk and allianceandleicester.co.uk have different distributions

Experimental Setting

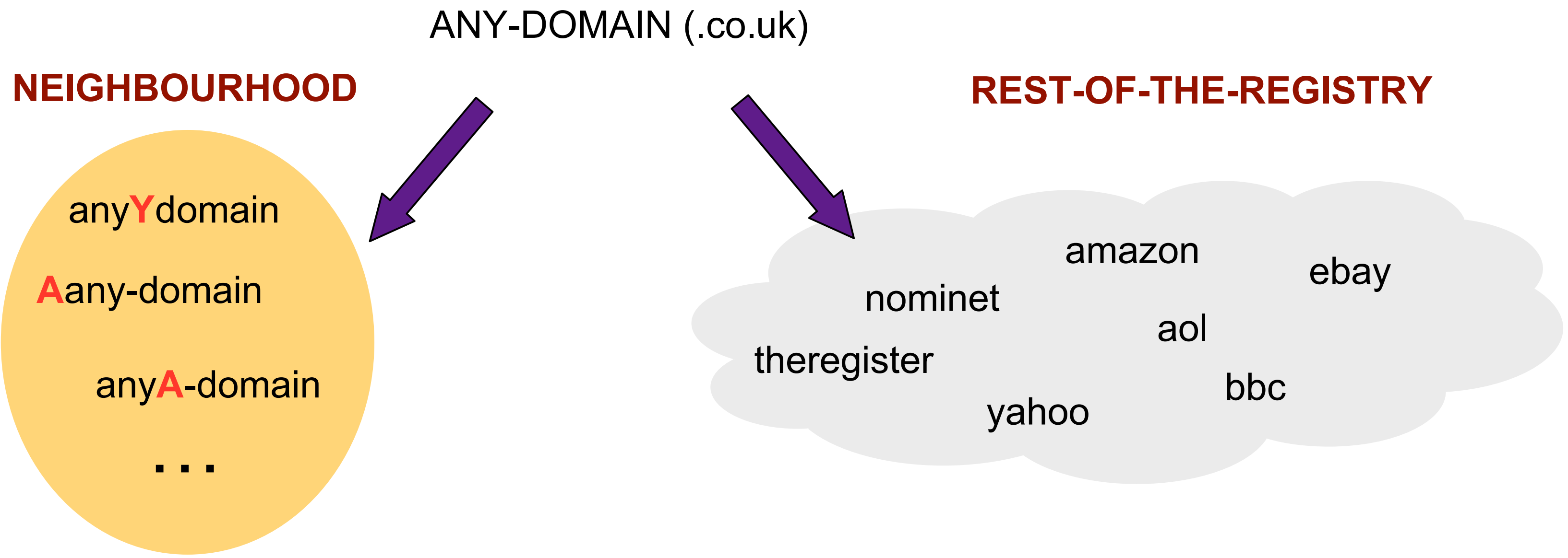
- Only .co.uk web sites considered (March 2008)
  - Length refers to the third-level label
- Set of random domains (expected behaviour)
  - 1000 domains for each length (random sample)
- Set of top-1000 popular domains (source NetCraft.com)
  - Band A: domains with ranking in [1,100]
  - Band B: domains with ranking in [101,500]
  - Band C: domains with ranking in [501,1000]

Neighbourhood and Popularity



nominet

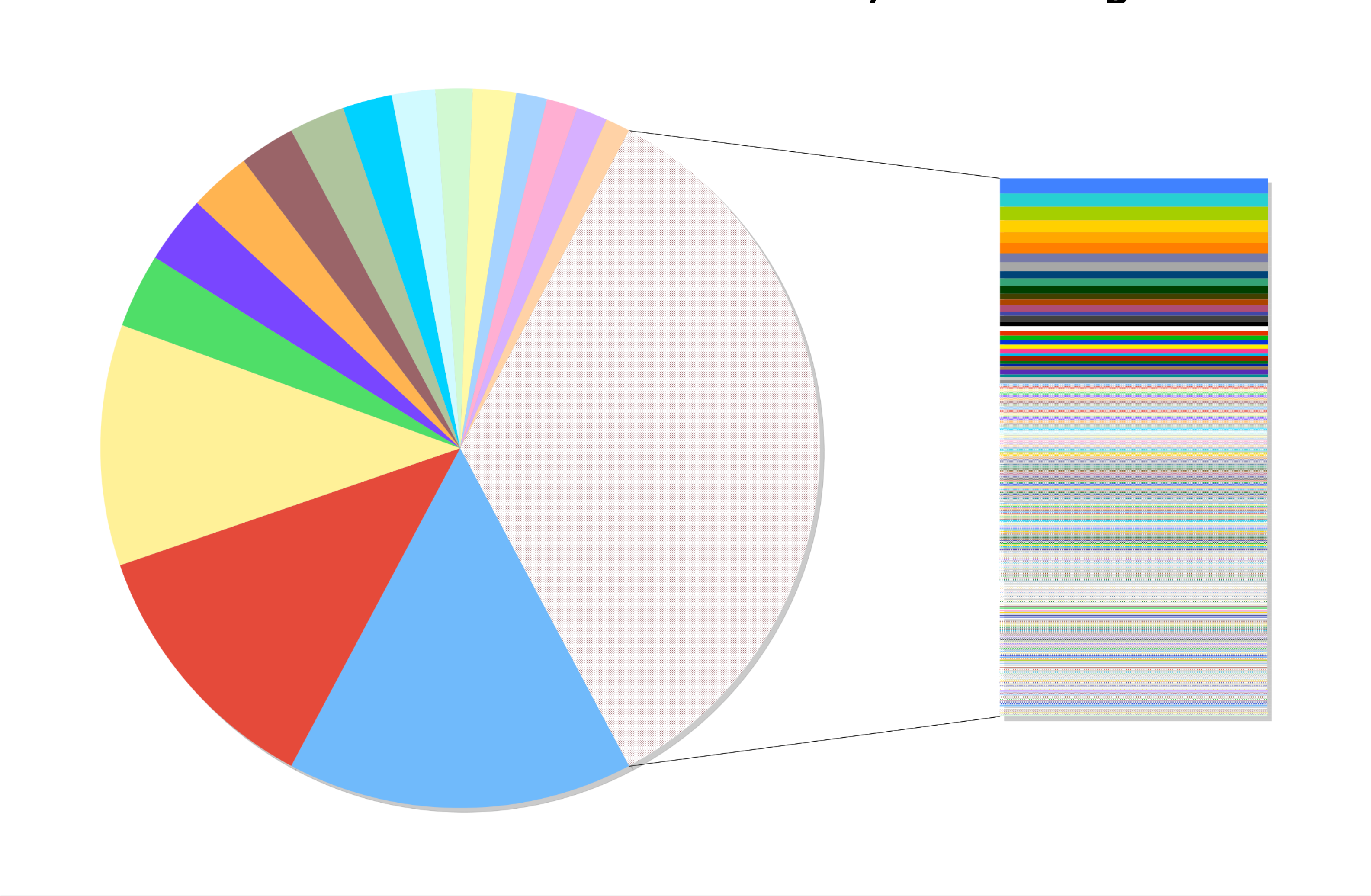
Outline



- Correlation between popularity of a domain name and size of its neighbourhood
- Presence of "typosquatters friendly" registrars in the neighbourhood of popular domains

Distribution of Registrars

- Fraction of domain names owned by each registrar

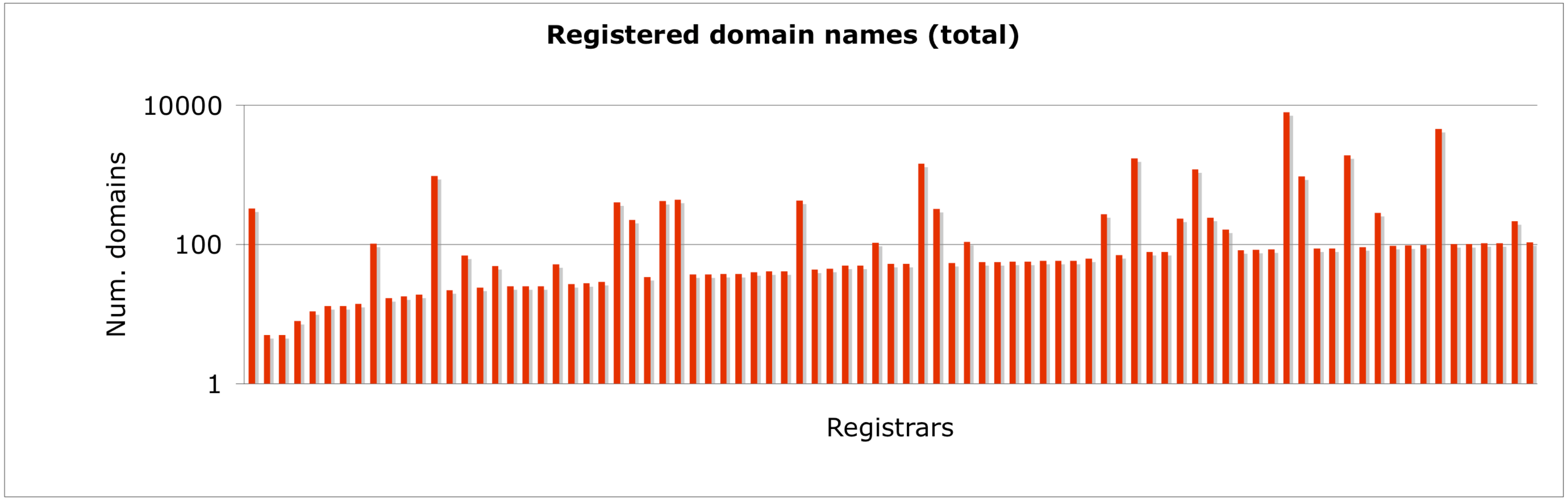


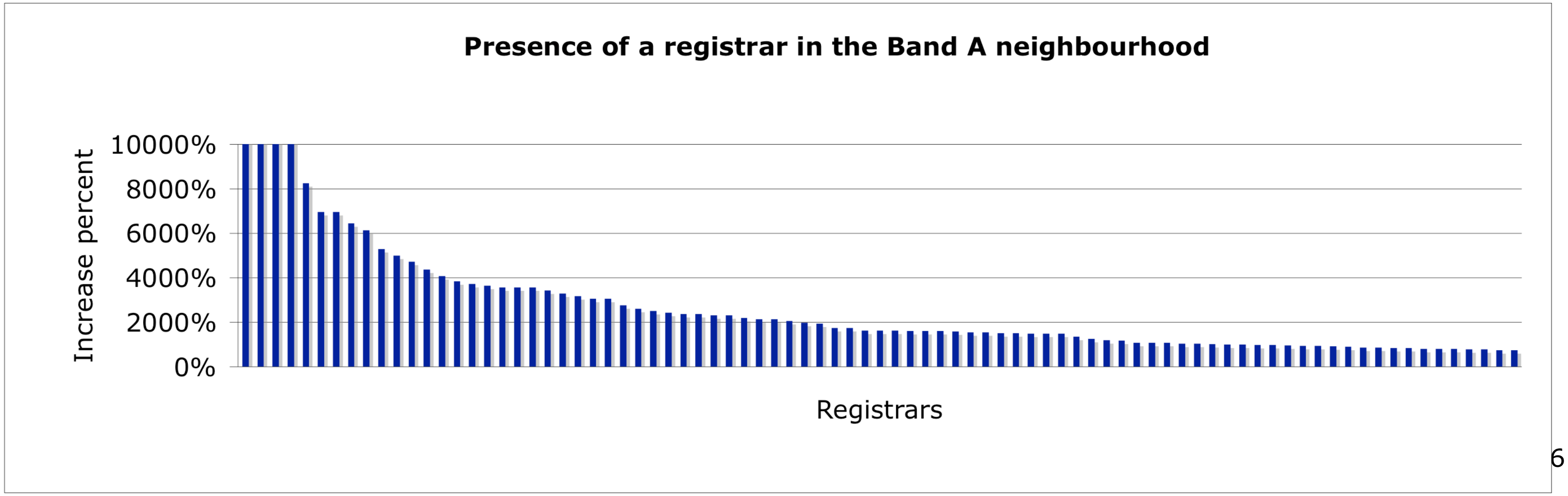
Experimental Setting

- Consider only domains in Band A’s neighbourhood
  - i.e., any domain at dist=1 from at least one domain in Band A
- Compute the number of domains owned by each of registrars (distribution)
- For each registrar, compute the percent increase wrt to the previous distribution

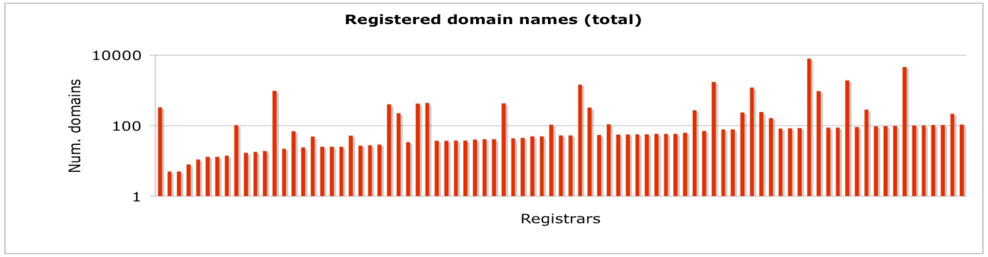
$$I\% = \frac{FracDom(BandA) - FracDom(registry)}{FracDom(regisry)} \cdot 100$$

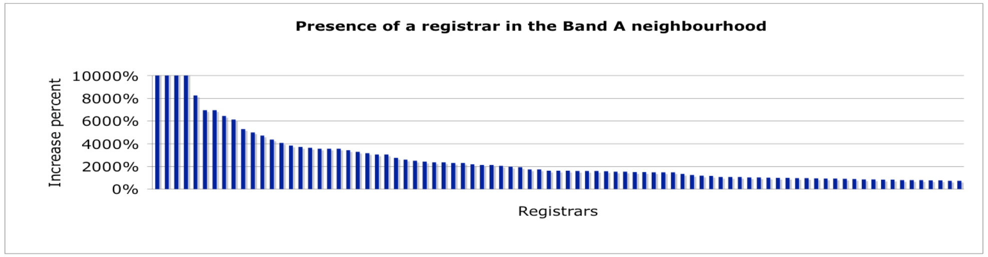
Distribution of Registrars (Band A)





Discussion





- Analysis (manual) of 25 registrars whose size is between 100 and 1000 domains
  - Big registrars are complex to analyse (not present in this chart)
  - Small registrars do not contribute to reliable statistics

Discussion

- One of the big domain names owns the majority of its neighbourhood
- Interesting activity for ~~6~~ 5 registrars
  - A big fraction of their domains syntactically or confusingly similar to popular domain names
- Normal activity for 8 registrars (false positives)
- No relevant findings in the other cases

Further Research Directions

- Insight in the typosquatting phenomenon
  - Domain name neighbourhood
  - First attempt toward statistical characterisation
- More questions than answers
  - Name servers used by typosquatters
  - Domain names containing common words
  - Content of the website
  - …

Bibliography

A. Banerjee, D. Barman, M. Faloutsos, Laxmi Bhuyan. *Cyber-Fraud is One Typo Away*. INFOCOM, 2008

Y.Wang, D. Beck, J. Wang, C. Verbowski, B. Daniels. *Strider Typo-Patrol: Discovery and Analysis of Systematic Typo-Squatting*. SRUTI (Usenix WS), 2006.

McAfee. *What’s In A Name: The State of Typo-Squatting 2007*. http://us.mcafee.com/root/identitytheft.asp?id=safe_typo (valid as in Jan. 2008).

WIPO. *DNS Developments Feed Growing Cybersquatting Concerns*. http://www.wipo.int/pressroom/en/articles/2008/article_0015.html (valid as in May 2008).

nominet

Thank you!!!

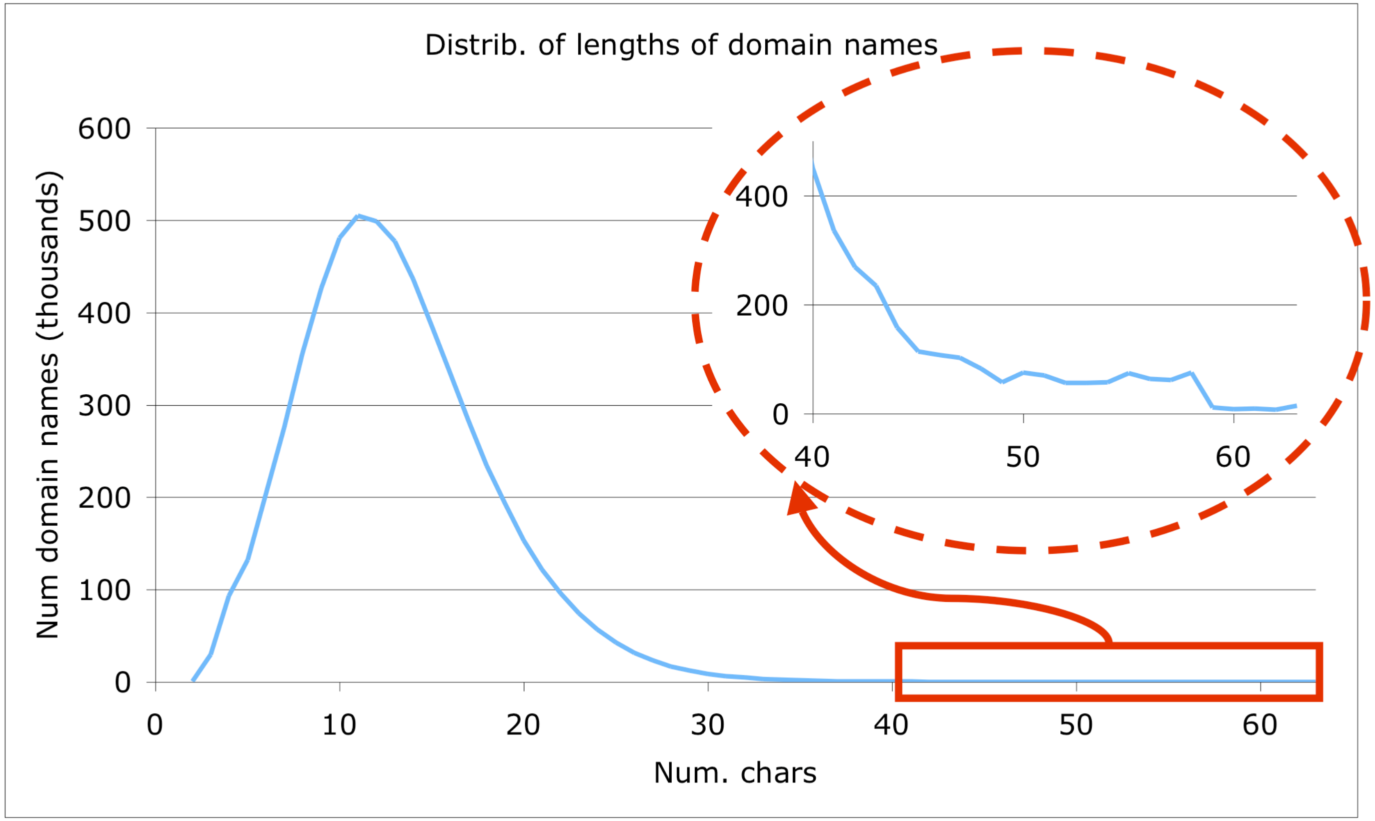
nominet

Questions?

nominet

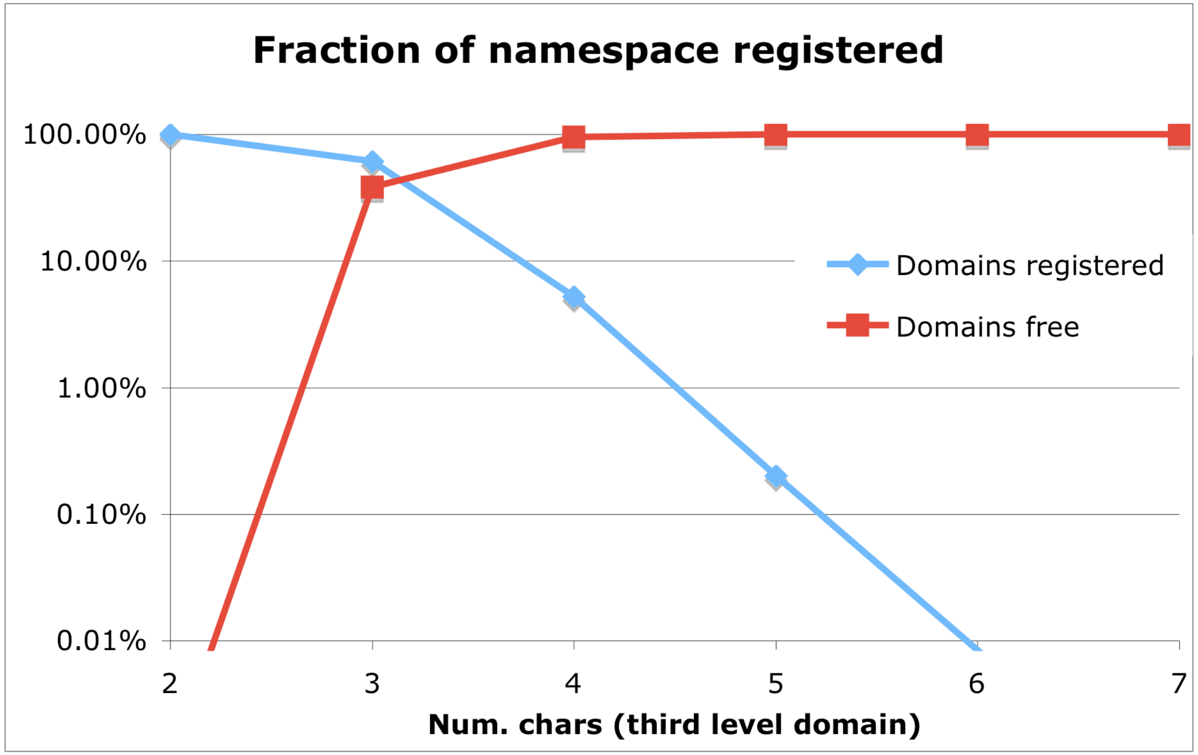
Backup Slides

Length of a domain name



- co.uk domains only
- Length always refers to the third level domain

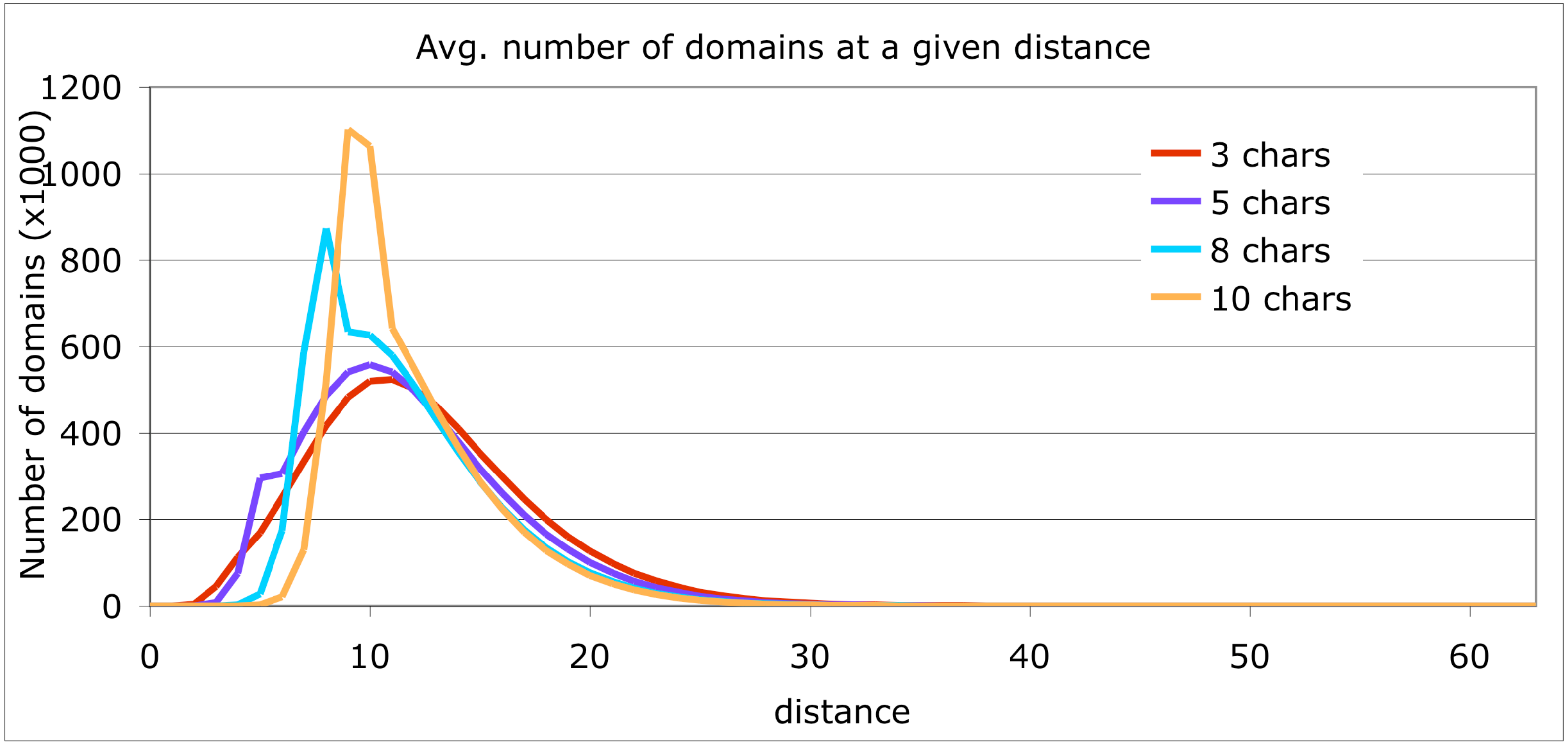
Length of a domain name



- 3- and 4- chars domains not meaningful
- Neighbourhood of 5-chars domains is in the 4-chars space

Statistical characterisation

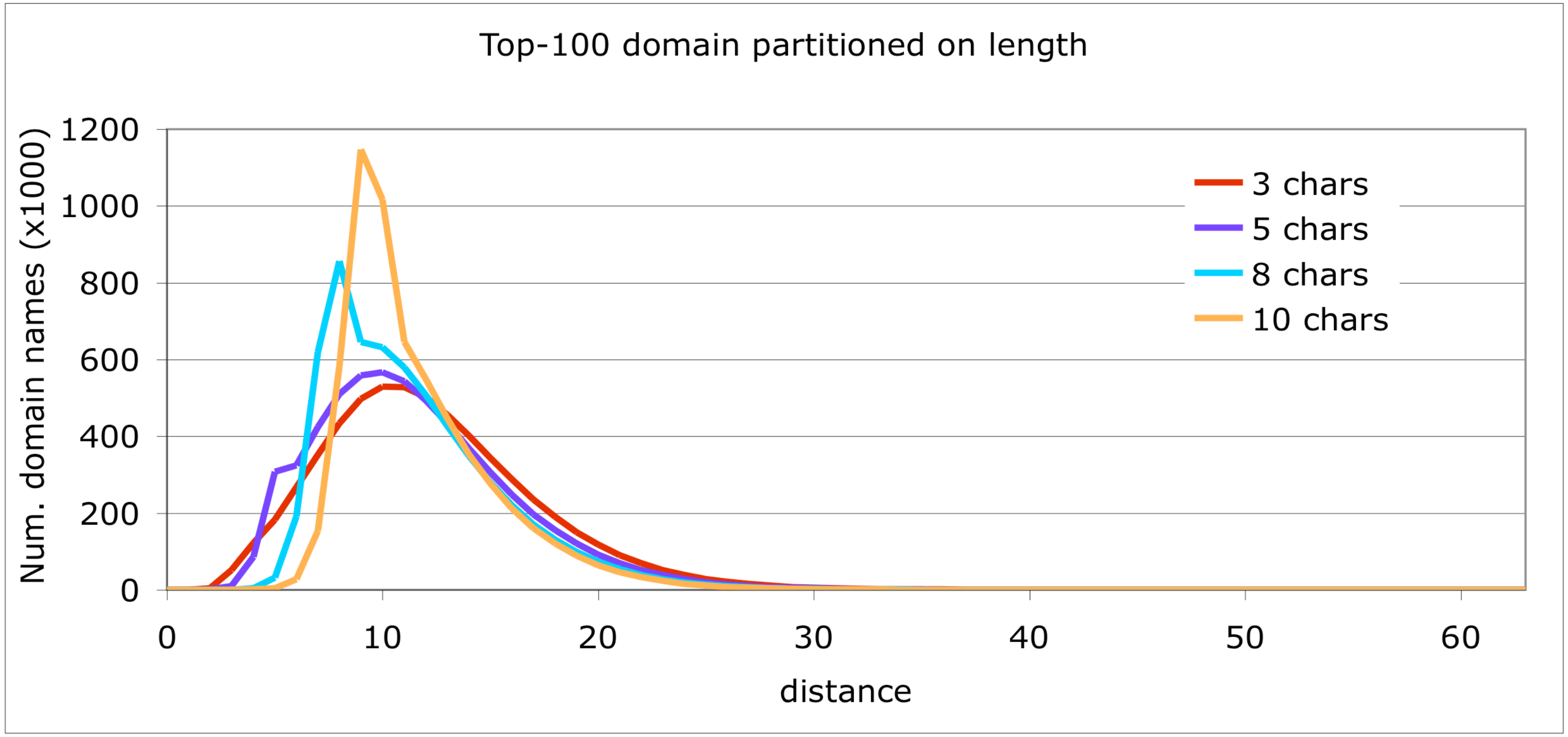
Distance between domain names



- ~100 domains (for each length) compared against whole dataset
- Average number of domains at a given distance

Statistical characterisation

Top-100 (band A) domain names



- ~10 domains (for each length) compared against whole dataset
- Average number of domains at a given distance