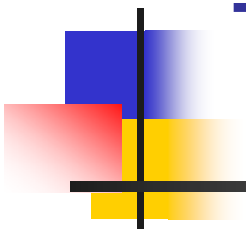


The Spanish Group

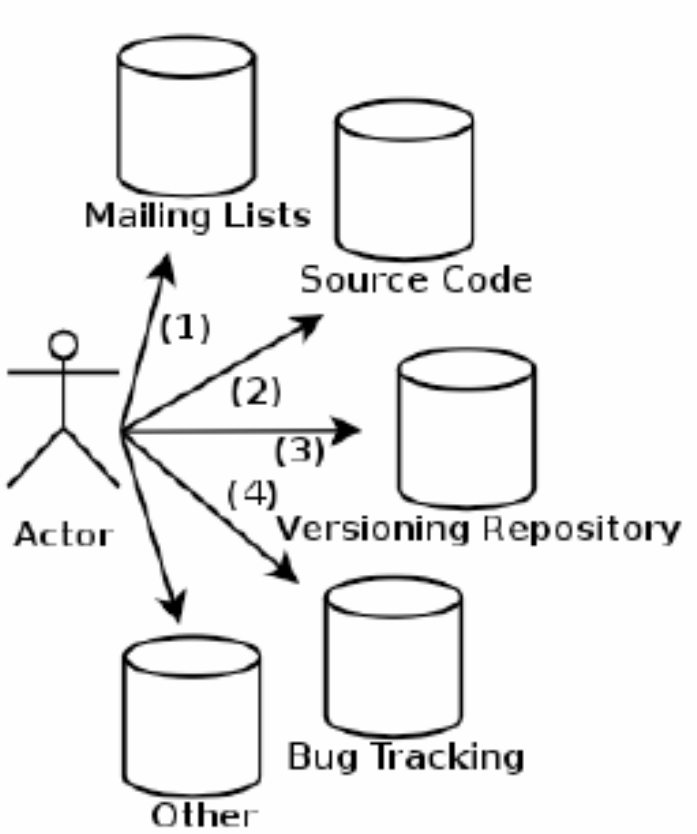
Presented Peter Rigby

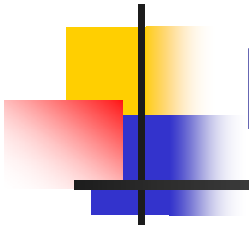
MSR Class

November 2, 2006

Developer Identification

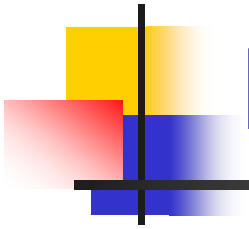
- Identify individuals across multiple data sources
- May also have multiple identities within a single source





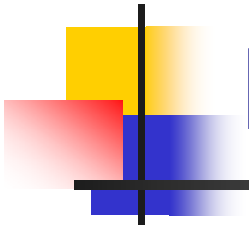
Heuristics

- Link secondary and primary info
- GPG key
- Email address, e.g., pcr@uvic.ca
- Match cvs user name with email
- Module owner file or cvs write file
- All very specific and error prone
  - Need some kind of metric to assess error
  - Manual intervention required



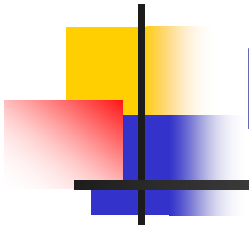
Example of data they sent me

- **dean gaudet's identity: (there is at least one other dean on the project)**
- dean@arctic.org; dean gaudet (output from Gregorio)
- **Ones that I feel should have been automatically added: (MANUAL)**
- dgaudet@iacnet.com
- dgaudet@arctic.org
- dean-list-new-httpd@arctic.org
- dgaudet@hotwired.com
- dgaudet@wired.com
- dgaudet@hyperreal.org
- dgaudet-list-linux-kernel@arctic.org
- **Ones that are ambigious and should be flagged for manual resolution (searching for gaudet reveled no ambiguities): (MANUAL)**
- dean@go.co.uk
- dean@myp.com



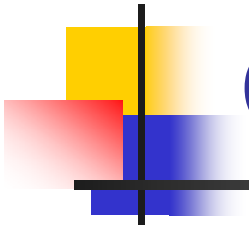
Bird et al. MSR 2006

- Normalize names
- Levenshtein edit distance
  - First and last names
- Names-email similarity
- Email base similarity
- Cumulative ID similarity
- Creates large clusters, manually split



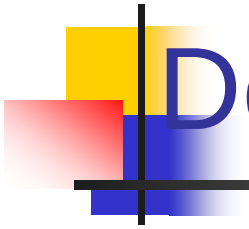
Privacy issues

- Technical description of how to link together identities (hash)
- Firewall?
  - The identification table keeps ids safe
- “[someone] can always milk the same repositories, and obtain exactly the same data”



Conclusions

- Strength
  - Identifying individuals is difficult and they provide some useful ideas for doing this
  - Bird provides an interesting approach
- Weakness
  - No error rate or assessment of accuracy
  - Db plan not meaningful without good heuristics
  - Did not perform well on single Apache mailing list

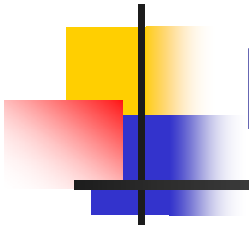


Developer Geographic Location

- Want to understand where developers are from
- Previous work:
  - Europe taking over from US on Debian and Linux (credit file)
- Examined SourceForge data
  - Private emails addresses and time zone

Methodology

- Email country code (e.g., “.ca”)
  - “.com” especially in USA
- Time zones
  - Often not country specific (e.g., PST)
  - “it is trivial to assign a time zone to a country” e.g., EST = us!? What about Mexico or Ecuador or Canada?



Methodology (Cont.)

- Not indicative of how much the user actually participates
- I have 2 SourceForge accounts, but rarely use SourceForge

GMT

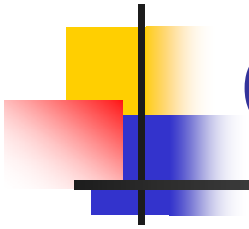
- Find ratio of users with identifable time zone and address vs only address or only time zone
- Redistribute GMT based on these ratios
- So if 20 hotmail.com are (EST) of a total of 60 then 1/3 of the GMT, hotmail.com users would be EST
- Uk, ie, and pt which are GMT use european ratio, because they are in GMT

Results

- Find that identifiable domains provides statistically similar outcomes to more complicated techniques
- Europe vs. NA

Region	Developers
Africa	12560
Asia	127275
EU	401845
Europe	466792
North America	485679
Oceania	46422
South America	36330

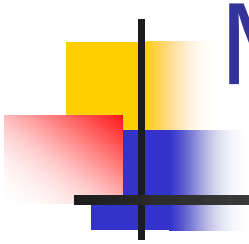
Table 7: Results by regions.



Conclusions

- Strengths
  - Large data set
  - Original approach with time zones and email addresses
  - Email address top domain is a useful predictor
- Weaknesses
  - Having does not indicate doing
  - Is .us less used than .de or .ca
  - Very rough estimates
    - (e.g.,) Time zones don’t divide evenly

My Approach



- Extract time zone from sent header
- On dev mailing list so indicates
  - Measure of activity, could also do it for individuals
- Results for apache 1995-2005
  - total = 104650, correct = 99684,
  - Error = 4966 -> 5%