Open access to “big data” for health research and the public good

Ingemar J. Cox
Introduction

• Part 1: A brief history of me
• Part 2: Open access to big data
The early years
AT&T Bell Labs

• Blanche: An experiment in
AT&T Bell Labs

- Simultaneous localisation and mapping (SLAM)
NEC Research Institute

- Stereo correspondence
- Motion correspondence
- Content-based image retrieval
- Digital watermarking
NEC Research Institute

- Stereo correspondence
NEC Research Institute

- Motion correspondence
NEC Research Institute

- PicHunter: Content-based image retrieval
NEC Research Institute

• Digital watermarking
roughly $\sqrt{n}/t$. Thus, the similarity measure can be shrunk by a factor of $t$.

We do not know of any more effective multidocument attack on normally distributed watermarks. In a forthcoming paper (see http://www.neci.nj.nec.com/tr/index.html), a more theoretical justification is given for why it is hard to achieve more than an $O(1)$ reduction in the similarity measure.

V. EXPERIMENTAL RESULTS

In order to evaluate the proposed watermarking scheme, we took the Bavarian couple image of Fig. 4 and produced the watermarked version of Fig. 5. We then subjected the watermarked image to a series of image processing and collusion style attacks. These experiments are preliminary, but show resilience to certain types of common processing. Of note is our method’s resistance to compression such as JPEG, and data conversion (printing, xeroxing and scanning). Note that in the case of affine transforms, registration to the original image is crucial to successful extraction.

In all experiments, a watermark length of 1000 was used. We added the watermark to the image by modifying 1000 of the more perceptually significant components of the image spectrum using (2). More specifically, the 1000 largest coefficients of the DCT (excluding the DC term) were used. A fixed scale factor of 0.1 was used throughout.

A. Experiment 1: Uniqueness of Watermark

Fig. 6 shows the response of the watermark detector to 1000 randomly generated watermarks of which only one matches the watermark present in Fig. 5. The positive response due to the correct watermark is very much stronger than the response to incorrect watermarks, suggesting that the algorithm has very low false positive response rates.

B. Experiment 2: Image Scaling

We scaled the watermarked image to half of its original size, as shown in Fig. 7(a). In order to recover the watermark, the quarter-sized image was rescaled to its original dimensions, as shown in Fig. 7(b), in which it is clear that considerable fine detail has been lost in the scaling process. This is to be expected since subsampling of the image requires a lowpass spatial filtering operation. The response of the watermark detector to the original watermarked image of Fig. 5 was 32.0, which compares to a response of 13.4 for the rescaled version of Fig. 7(b). While the detector response is down by over 50%, the response is still well above random chance.

The common test image “Lenna” was originally used in our experiments and similar results were obtained. However, Playboy Inc. refused to grant copyright permission for electronic distribution.
Signafy

• First spinout company from NEC Research Institute
• Co-founder and CTO
University College London (UCL)

- Adastral Park
- Information retrieval
- Early warning sensing systems for infectious diseases
UCL

• Adastral Park
UCL

- Information retrieval
- Efficient evaluation of IR systems
- P2P information retrieval
- P2P Twitter
UCL

- EPSRC IRC Early Warning Sensing Systems for Infectious Diseases
Part 2: Open access to big data
How big is big?
Big data: examples

- Google
  - 115B queries per month
  - 1.2 billion unique users
- Facebook
  - 1.1 billion unique users per month
- Twitter
  - 218 million active users
Never has so much been known of so many by so few
Unexpected benefits

- Flu trends
- Phase 4 trials
- Anorexia
- Mass gatherings
Phase 4 trials

Studies conducted after a drug has been approved, to determine additional information about the drug's risks, benefits, and optimal use.
Phase 4 trials


Phase 4 trials

176 million unique users were included in the study
Anorexia

- Flickr
- Pro-anorexics
  - thinspiration
- Recovering-anorexics
Mass gatherings

- Music festivals, sports events, religious gatherings
- Possible spread of infectious disease
- Use Twitter and search logs for surveillance
Nuria Oliver: what big data and the Mexican pandemic taught us

We need to harness the power of anonymised big data and do social good, Nuria Oliver told the audience at Wired 2013, imploring public and private companies to come together to achieve this.

Oliver led a team at Telefonica Digital, where she works as scientific director of its research and development department, to prove the worth of this statement in the aftermath of the H1N1 flu outbreak in Mexico in 2009.
What value to society?

Who knows.
What value to society?

**EUROPEAN COMMISSION - PRESS RELEASE**

**Digital Agenda: Turning government data into gold**

Brussels, 12 December 2011 – The Commission has launched an Open Data Strategy for Europe, which is expected to deliver a €40 billion boost to the EU’s economy each year. Europe’s public administrations are sitting on a goldmine of unrealised economic potential: the large volumes of information collected by numerous public authorities and services. Member States such as the United Kingdom and France are already demonstrating this value. The strategy to lift performance EU-wide is three-fold: firstly the Commission will lead by example, opening its vaults of information to the public for free through a new data portal. Secondly, a level playing field for open data across the EU will be established. Finally, these new measures are backed by the €100 million which will be granted in 2011-2013 to fund research into improved data-handling technologies.

These actions position the EU as the global leader in the re-use of public sector information. They will boost the thriving industry that turns raw data into the material that hundreds of millions of ICT users depend on, for example smart phone apps, such as maps, real-time traffic and weather information, price comparison tools and more. Other leading beneficiaries will include journalists and academics.

Commission Vice President Neelie Kroes said: “We are sending a strong signal to administrations today. Your data is worth more if you give it away. So start releasing it now: use this framework to join the other smart leaders who are already gaining from embracing open data. Taxpayers have already paid for this information, the least we can do is give it back to those who want to use it in new ways that help people and create jobs and growth.” [See Mrs Kroes video quote here](#).

The Commission proposes to update the 2003 Directive on the re-use of public sector information by:

- Making it a general rule that all documents made accessible by public sector bodies can be re-used for any purpose, commercial or non-commercial, unless protected by third party copyright;
- Establishing the principle that public bodies should not be allowed to charge more than costs triggered by the individual request for data (marginal costs); in practice this means most data will be offered for free or virtually for free, unless duly justified.
- Making it compulsory to provide data in commonly-used, machine-readable formats, to ensure data can be effectively re-used.
- Introducing regulatory oversight to enforce these principles;
- Massively expanding the reach of the Directive to include libraries, museums and archives for the first time; the existing 2003 rules will apply to data from such institutions.
2. **Open data, why does it matter for Europe?**

2.1. **Untapped business and economic opportunities**

The 2009 Digital Britain Report described data as ‘an innovation currency’ and ‘the lifeblood of the knowledge economy’\(^2\). A large part of this innovation currency is produced, collected or paid for by governments across the EU. It is an essential raw material for a wide range of new information products and services that build on new possibilities to analyse and visualise data from different sources.

These products range from decision support systems for businesses, location-based services and car navigation systems to weather forecasts and other ‘apps’ for our smartphones.

The market size and growth of the geographic information sector shows the potential of public data as an engine for job creation. The German market for geo-information in 2007 was estimated at €1.4 billion, a 50% increase since 2000\(^3\). In the Netherlands, the geo-sector accounted for 15,000 full time employees in 2008. Other areas such as meteorological data, legal information and business information also form the basis of steadily growing markets.

A recent study estimates the total market for public sector information in 2008 at €28 billion across the EU\(^4\). The same study indicates that the overall economic gains from further opening up public sector information by allowing easy access are around €40 billion a year for the EU27. The total direct and indirect economic gains from PSI applications and use across the whole EU27 economy would be in the order of €140 billion annually.

Data management is also essential for tackling environmental challenges. Examples are the processing of energy consumption patterns to improve energy efficiency or of pollution data in traffic management. Informed policy decisions in the areas of transport, land use and climate change depend increasingly on analysis of the available data.

---


\(^3\) Assessment of the Re-use of Public Sector Information in the Geographical Information, Meteorological Information and Legal Information sectors, MICUS, December 2008.

\(^4\) Review of recent studies on PSI re-use and related market developments, G. Vickery, August 2011.

Why don’t companies make their data available?

- In August 2006 AOL researchers published a sample query log from a 3 month period
  - 20M queries
  - 658K unique users
  - This represents 0.33% of all queries during this period
Why don’t companies make their data available?

• The log consists of:
  • UserID,
  • Query,
  • QueryTime,
  • ClickedRank,
  • DestinationDomainUrl
<table>
<thead>
<tr>
<th>AnonID</th>
<th>Query</th>
<th>Query time</th>
<th>ItemRank</th>
<th>clickURL</th>
</tr>
</thead>
<tbody>
<tr>
<td>8760</td>
<td>jojo the singer</td>
<td>2006-03-26 16:02:04</td>
<td>5</td>
<td><a href="http://www.jojofan.com">http://www.jojofan.com</a></td>
</tr>
<tr>
<td>8760</td>
<td>jeniffer lopez</td>
<td>2006-03-26 16:05:29</td>
<td>4</td>
<td><a href="http://www.allstarz.org">http://www.allstarz.org</a></td>
</tr>
<tr>
<td>8760</td>
<td>jeniffer lopez</td>
<td>2006-03-26 16:05:29</td>
<td>10</td>
<td><a href="http://www.starpulse.com">http://www.starpulse.com</a></td>
</tr>
<tr>
<td>8760</td>
<td>nicole richie</td>
<td>2006-03-26 17:28:58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8760</td>
<td>free porn</td>
<td>2006-03-28 16:43:16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Why don’t companies make their data available?

• AOL dismissed Abdur Chowdhury, the researcher who posted the data, along with another employee.

• Maureen Govern, AOL’s chief technology officer, resigned.

• The EFF and World Privacy Forum filed complaints with the FTC.
No good deed goes unpunished.
What should be done?

• No clear answer
• But current debate only focuses on privacy
• The question is not so simple
What should be done?

• Technology
• Legal
  • limited liability?
• Ownership of data
  • individuals or organisations?
Thank you.
Comments and suggestions welcome.