From Panopticon to Fresnel Dispelling A False Sense of Security

Jon Crowcroft,
http://www.cl.cam.ac.uk/~jac22
With Ian Brown, OII
Introduction

- Marconi Prof at Cambridge Computer Laboratory
- Spent 22 years at University College London (CS)
  - Home of Jeremy Bentham (creator of panopticon)
  - Presented my first paper here @ UKC in 1982 (t - 30 years:)
- Topic - ubiquity of sensors & privacy
- Sub topics -
  - differential privacy - limitations
  - Privacy of graph data - limitations
  - privacy by design
- Thanks to Siani Pearson of HP for Monday's Keynote
  - provides perfect background, tutorial & definitions!
What are you doing?
Where are you now
Bringing it all back home

Total Situational Awareness
Federating Sensor Nets

- Fresnel is an EPSRC funded project
- To Federate Sensor Nets
- But provide isolation and privacy
- And tools (intellectual and technical) for
- *Privacy by design*
Isolation
Virtualisation

(a)

(b)

Hop distances

\( a < b \)

\( a < p \)
Fresnel App
Deployment

- 34 Nodes (29 permanent)
- 10 Rooms
- iMote2 Sensors attached to desks
- Two applications
  - Desk occupancy
  - Environmental monitoring
Deployment Results
Fresnel Use

We can see you
Fresnel Worry

Pan-panopticon

![Graph showing relationship between arrival time and privacy concern.](image)

- $r^2 = 0.670$
- $P < 0.0001$
• Why am I telling you about this?

• Be afraid...you are’nt just under the fresnel lens

• It’s a Pan-panopticon...be very afraid :-(

Take Homes
Aggregation of Sources

problem

• But what about the data, once it's bought back to the cloud?

• Well, then we need different tools - which is what we talk about next...
Through a Graph, Darkly

- A Manifesto for Personal Privacy
- But still allowing data driven research
- And evidence based policy
- And all that targeted advertising:-)
- N.B. Of particular concern in UK right now, as
  1. Open Data Initiatives (including NHS)
  2. ICO is pushing for naïve anonymisation!!
  3. Communications Capability Dev. Programme
     Aka Communications Data Bill is a v. bad idea ... ... ...
Sharing Data

• We’re increasingly asked to share data
  • Government Open Data
  • Nature June article
  • EPSRC (&NIH&etc) funding

• We’re increasingly asked to be ethical
  • See Neuhaus&Webmoor “agile ethics for massified research and visualization”
  • Informed consent?
  • Anonymization/Data Privacy?

• Social Media very tempting study
  • Graph data & Personal Identifying Information(PII)
Complexity & Value

- Do people really understand?
  - Recent court cases on Terms & Conditions
  - Judgement was no (spread betting case)
  - 40 pages of legalise is not comprehensible
  - And therefore not valid

- Do people know the value of their PII?
  - Well, there’s some disagreement:-
    - Preibusch&Jentzsch/Harasser: Monetizing Privacy
    - Yes, but its pretty cheap
    - Brown et al
    - No, but then when they lose it it’s very expensive
But graph data is so interesting...

- Social Network is incredibly tempting
  - Twitter, Fb, WWW, Co-authorship
  - Social Science, Medicine, Commercial Motives....

- Characteristics like degree distribution, assortativity, betweenness very useful:
  - Info flow (percolation, gossip, epidemic)
  - Viral marketing (opinion dynamics)
  - Attack/defense/immunise/quarantine

- Can we anonymize
  - Bonneau “8 friends are enough” say naïve is no use
  - Backstrom “Wherefore Art Thou R3579X” say even quite subtle, no
  - Sala “Sharing Graphs using Differentially Private Graph Models”, IMC 2011, say, finally, yes, but take great care only _model_
Nodes: people, Links: relations

- Looking at an abstract graph hides reality
- Node data is PII
- It's personal
- But collection of edge/link data can be used to identify nodes
- Even if PII is protected
Anonymizing node data

- If data is separate from graph, then anonymization is feasible.
- Risk of re-identification of records if not careful statistically
- Differential Privacy...
Differential Piracy example

- Imagine we have a database of pirates.
- If we query for a very tall pirate with a long beard, we are asking to identify a unique record ("Long John Silver")
- If we ask "How many pirates in Penzance?" we are safe, as there are lots
- Or if we ask for the number of 1 legged pirates who also have parrots?
- But don’t ask for the pirate with the prosthetic hand, coz that even tells you his name...
### Piracy Preserving Database

<table>
<thead>
<tr>
<th>Name</th>
<th>Port</th>
<th>Parrot</th>
<th>Wooden Leg</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>Penzance</td>
<td>y</td>
<td>y</td>
<td>1.75</td>
</tr>
<tr>
<td>y</td>
<td>Penzance</td>
<td>y</td>
<td>y</td>
<td>1.74</td>
</tr>
<tr>
<td>z</td>
<td>Penzance</td>
<td>y</td>
<td>y</td>
<td>1.76</td>
</tr>
<tr>
<td>Dread pirate roberts</td>
<td>?</td>
<td>n</td>
<td>n</td>
<td>1.80</td>
</tr>
<tr>
<td>Hook</td>
<td>Neverland</td>
<td></td>
<td></td>
<td>1.65</td>
</tr>
<tr>
<td>Shakespeare</td>
<td>Airport</td>
<td></td>
<td></td>
<td>1.60</td>
</tr>
<tr>
<td>Sparrow</td>
<td>Hollywood</td>
<td></td>
<td></td>
<td>1.50</td>
</tr>
<tr>
<td>Long john silver</td>
<td>Treasure island</td>
<td>y</td>
<td>y</td>
<td>2.00</td>
</tr>
</tbody>
</table>
## Piracy Preserving DBase

<table>
<thead>
<tr>
<th>#name</th>
<th>port</th>
<th>Parrot</th>
<th>Wooden leg</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>xxx</td>
<td>penzance</td>
<td>y</td>
<td>y</td>
<td>1.75</td>
</tr>
<tr>
<td>yyy</td>
<td>penzance</td>
<td>y</td>
<td>y</td>
<td>1.74</td>
</tr>
<tr>
<td>zzz</td>
<td>penzance</td>
<td>y</td>
<td>y</td>
<td>1.76</td>
</tr>
<tr>
<td>Dread pirate roberts (*)</td>
<td>?</td>
<td>n</td>
<td>n</td>
<td>1.80</td>
</tr>
<tr>
<td>foo</td>
<td>neverland</td>
<td></td>
<td></td>
<td>1.65</td>
</tr>
<tr>
<td>bar</td>
<td>airport</td>
<td></td>
<td></td>
<td>1.60</td>
</tr>
<tr>
<td>baz</td>
<td>hollywood</td>
<td></td>
<td></td>
<td>1.50</td>
</tr>
<tr>
<td>fie</td>
<td>Treasure island</td>
<td>y</td>
<td>y</td>
<td>2.00</td>
</tr>
</tbody>
</table>
Adding graph edges messes this

- Link data represents a lot of attacks on hash of name:
K-clique analysis reveals...
Minor External Knowledge.

• If I know that graph is of 2 years of undergrads,
• And the names of the class reps for y1&y2...
• I can infer those nodes identity trivially
• In the intersection of two cliques
• (the class reps meet together once per week)
• If this was student health record,
• I would have re-identified those two students... ...
Lots more graph properties...

- Degree of nodes
- All the centrality types (including spectral etc)
- If links have properties too (strength, as in recommendation or reputation, or age, or other)
- Worse than ever!
Worse to come...

- Dunbar’s # - 150
  - So if friend id is 32 bits, your friend list is 4800 bits on average
  - So the attack surface for identifying you is huge
- Worse Still - you have lots of “edges”
You have an edge for each type of relationship

- kin, friend, colleague
- Co-author of work
- Co-located (e.g. paid congestion charge same time, used oyster card on same journey, checked in on foursquare same place)
- Pay tax together, live at same postcode,
- Sent SMS, IM, Email, Phone call, cell phone call from location
- Same smart meter address
Re-identification is trivial

- Anyone in possession of 8 (see Anderson et al) I-Ds a graph of one set of edge type, with access to “anonymized” any other graph edge types, can re-identify the whole thing
- E.g. Tesco’s clubcard can re-identify your whole health net.....
Dynamics

- Forgetting might help reduce attack surface
- Remove edges from old (therefore less trusted)
Manifesto

- Separate storage of node PII and link data
- Always crypt PII
- Decentralize nodes and links
- Partition PII by role
  - Kin, friend, work, school
  - Health, finance, gov, social
- Use differential privacy on graph data as well as node data
- Make it easy to understand
  - Maybe add forgetting
Take homes

- Doesn’t have to be all central
  - Cannot guarantee safe way to share graphs (sorry:-(
  - Can use Differential Privacy for node data records (without graph)
  - Can do diff priv on graphs but need to take care on
  - multiple priv pres queries can still snowball

- Epidemiologists don’t need our bank data, government don’t need our social data, we don’t need your health statistics
Questions?

- Remembering I’ll be recording the Q&A
- And who you are:-)
The Fresnel Team...
There is a vision of a sensor rich world
Sensing available everywhere
Context-aware apps that use sensing infrastructure
Are we there?
Typical sensor networks are **fit-for-purpose**

- **Single-app**: Inflexible to updates and addition of services
- **Single-user**: High cost of deployment and maintenance
Shared Sensing Infrastructure
Management

- Over-the-air installation
In-node Virtualization

- Multiple applications co-live on the same node
• **Isolation:** Applications should operate as if they are running over a dedicated sensor node/network.

• **Overlay virtual sensor network** for each application
The SenShare Platform
In-node Architecture

Multi-tasking OS

Node Management

Network Manager

Sensing Scheduler

I/O API

Sensing
Net I/O
Clock

I/O API

Sensing
Net I/O
Clock

I/O API

Sensing
Net I/O
Clock

Timer
USB
RF
ADXL
Temp
Light
Accel
Humidity
SenShare

Multi-tasking OS

Node Management

Network Manager

Sensing Scheduler

I/O API

Sensing
Net I/O
Clock

Timer
USB
RF
ADXL
Temp
Light
Accel
Humidity

Network Manager

Sensing Scheduler

I/O API

Sensing
Net I/O
Clock

Timer
USB
RF
ADXL
Temp
Light
Accel
Humidity
Application Support

- **Support:**
  - TinyOS
  - C/C++
  - Future: Contiki

- Only the I/O is virtualized (not the binary)
• The SenShare API is integrated with the TOS platform
• Application’s source code does not require modifications if HIL is used
The SenShare API is integrated with the TOS platform
Application’s source code does not require modifications if HIL is used
C/C++ Application API

- Applications are linked with our library
- Hardware Independent API to:
  - Read/write to sensors
  - Send/receive network message
  - Interact with the runtime
Sensing Scheduler

• Virtualize I/O

• Advantages:
  • Common way to support all platforms
  • Access policies (privacy)
  • Throttle applications
  • Isolation

• Multiplexing
  • Asynchronous requests
  • Bursting queue
  • Serve all requests simultaneously
Virtualization Penalty
Node Management

- Allow the network owner and the users to
  - Configure Nodes
  - Change access policies
  - See what is running
  - Debug
  - Deploy/Install applications
  - Start/Stop applications
Network Support
## Selective Control

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOM_ID</td>
<td>Select a room</td>
</tr>
<tr>
<td>ROOM_TYPE</td>
<td>Select a room type (corridor, office, etc)</td>
</tr>
<tr>
<td>NODE_ID</td>
<td>Select individual nodes</td>
</tr>
<tr>
<td>NODE_TYPE</td>
<td>Select specific hardware (e.g., imote2)</td>
</tr>
<tr>
<td>SENSOR</td>
<td>Select specific sensor (e.g., temperature)</td>
</tr>
<tr>
<td>POWER_TYPE</td>
<td>Select on power type (battery, permanent)</td>
</tr>
<tr>
<td>AVAIL_POWER</td>
<td>Select on remaining power</td>
</tr>
<tr>
<td>NETWORK_LOAD</td>
<td>Select on average traffic</td>
</tr>
<tr>
<td>CPU_LOAD</td>
<td>Select on average CPU load</td>
</tr>
<tr>
<td>AVAIL_MEMORY</td>
<td>Select on available memory</td>
</tr>
<tr>
<td>AVAIL_STORAGE</td>
<td>Select on available storage</td>
</tr>
</tbody>
</table>

**Examples:**

- `<SELECT ALL; SENSOR=temp AND ROOM_TYPE=office AND NOT NODE_ID=5>`
- `<SELECT ONE; SENSOR=temp AND (ROOM_ID=1 OR ROOM_ID=2)>`
Selective Control
Selective Deployment and Installation

- Install new applications *over-the-air*
- Modified Deluge protocol*
- The CTP tree is used again

Participating nodes
Data Broadcast
Data Broadcast
Data Broadcast
Traffic Isolation

Hello

App 1  App 2  RUNTIME  Network

App 1  App 3  RUNTIME  Network
Traffic Isolation

Application header: \{app id, seq no, origin, destination\}
Overlay sensor networks

- Connect disconnected blobs
Overlay sensor networks

- Connect disconnected blobs
- Root over the CTP tree
Overlay sensor networks

- After the installation of a new application
  - Each node measures their distance to the closest cluster edge lower down the routing tree.
Message_received_from_node (A)

- if (local_distance == 0)
  - accept_message
- else if (distance(A) > (local_distance - 1))
  - send_to_closest_neighbour
- else
  - send_to_all_links(x) with distance(x) > distance(A)
  - send_to_parent
Other applications
Take away points

• We need to break the current, fit-for-just-one-purpose, model of SN.

• Allow users to install applications on shared infrastructure

• Provide the APIs, platform and tools that are required to support this.

• Do privacy right.
Future work

- Mobile phones
- Federation
- Privacy/security issues
- Economic model
Thanks!
Questions?

Download Link:
http://www.cl.cam.ac.uk/research/srg/netos/fresnel/