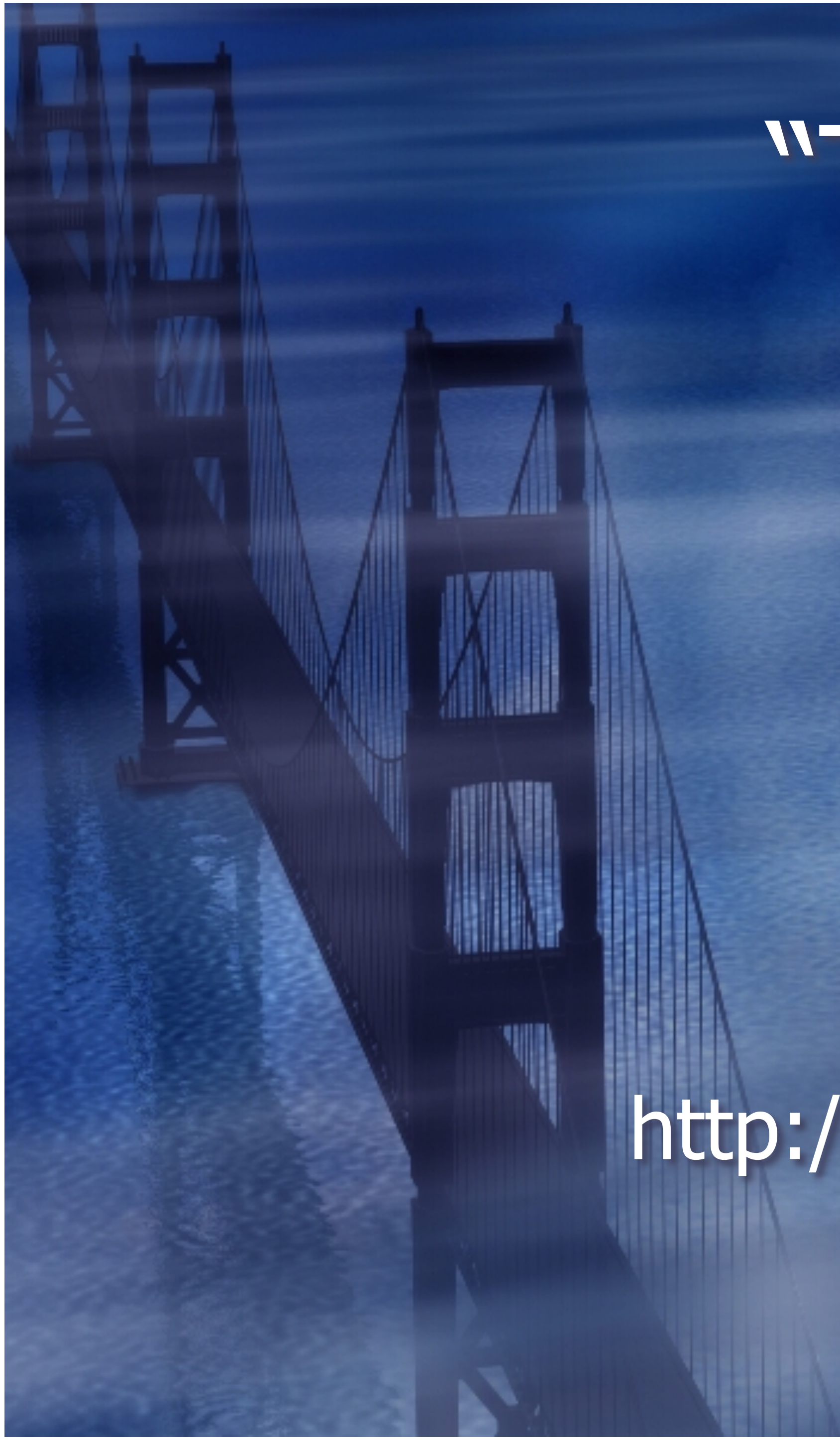
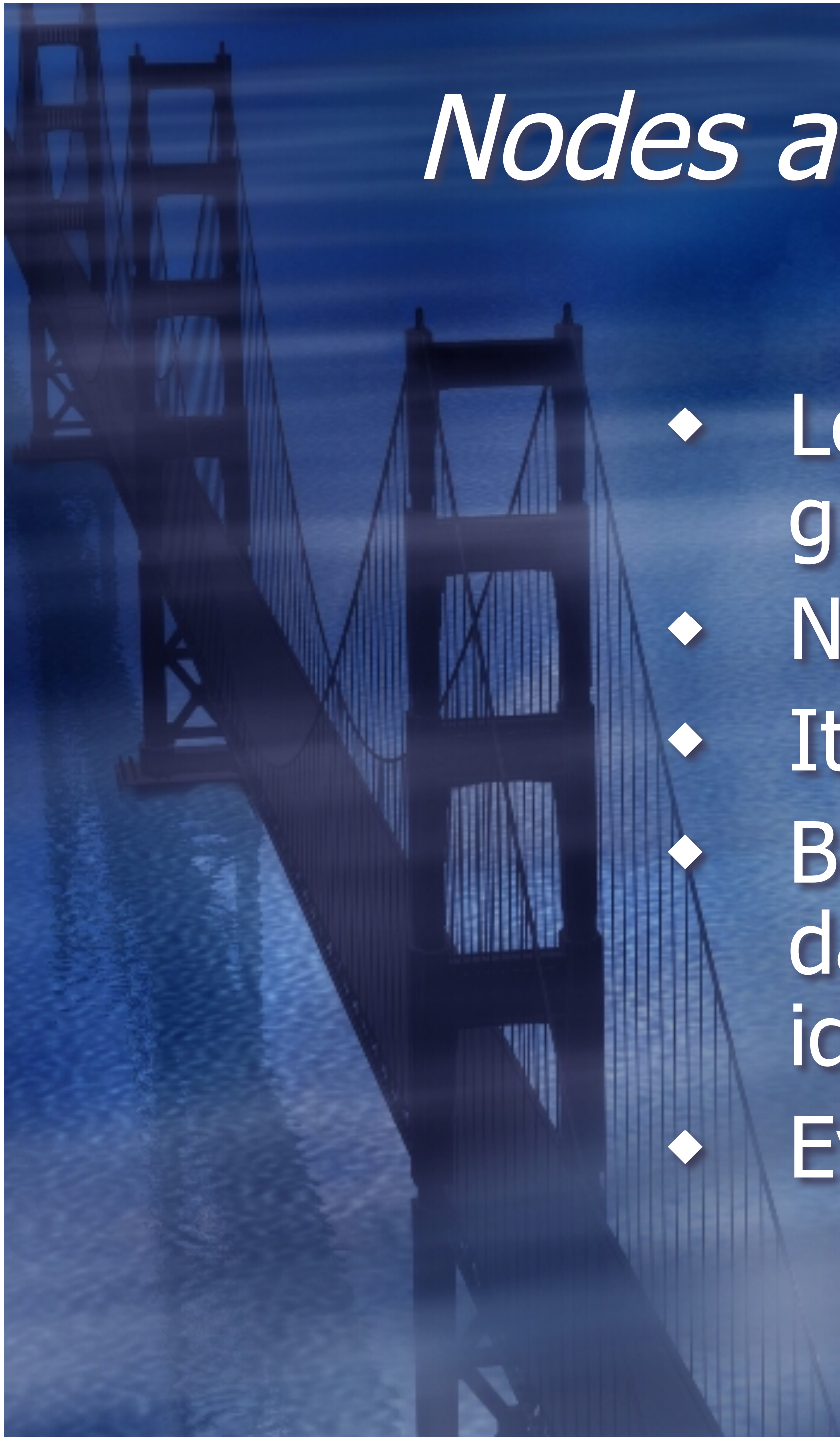


“Through a *Graph*, Darkly” or “A Manifesto for Online Privacy”

Jon Crowcroft,

http://www.cl.cam.ac.uk/~jac22

Nodes are people, links are relationships



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

Anonymizing Node Data Records

- 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 DBase

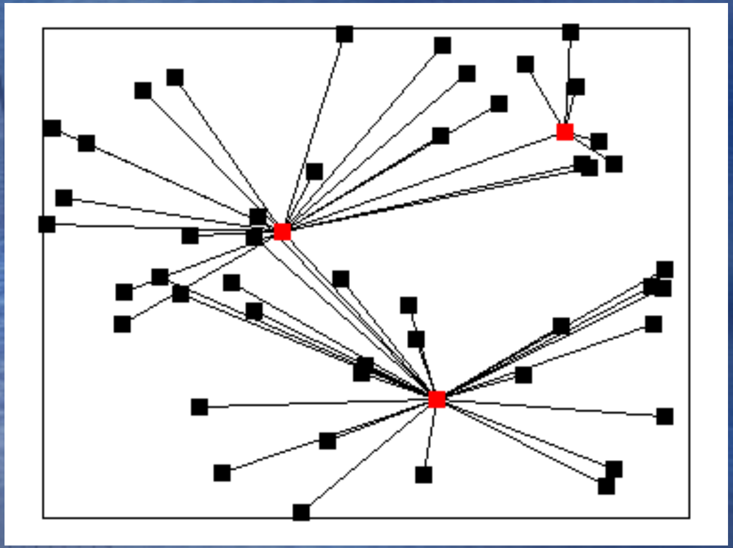
name	port	Parrot	Wooden leg	Height	
x	penzance	y	y	1.75	
y	penzance	y	y	1.74	
z	penzance	y	y	1.76	
Dread pirate roberts	?	n	n	1.80	
Hook	neverland			1.65	
shakespeare	airport			1.60	
sparrow	hollywood			1.50	
Long john silver	Treasure island	y	y	2.00	

Piracy Preserving DBase

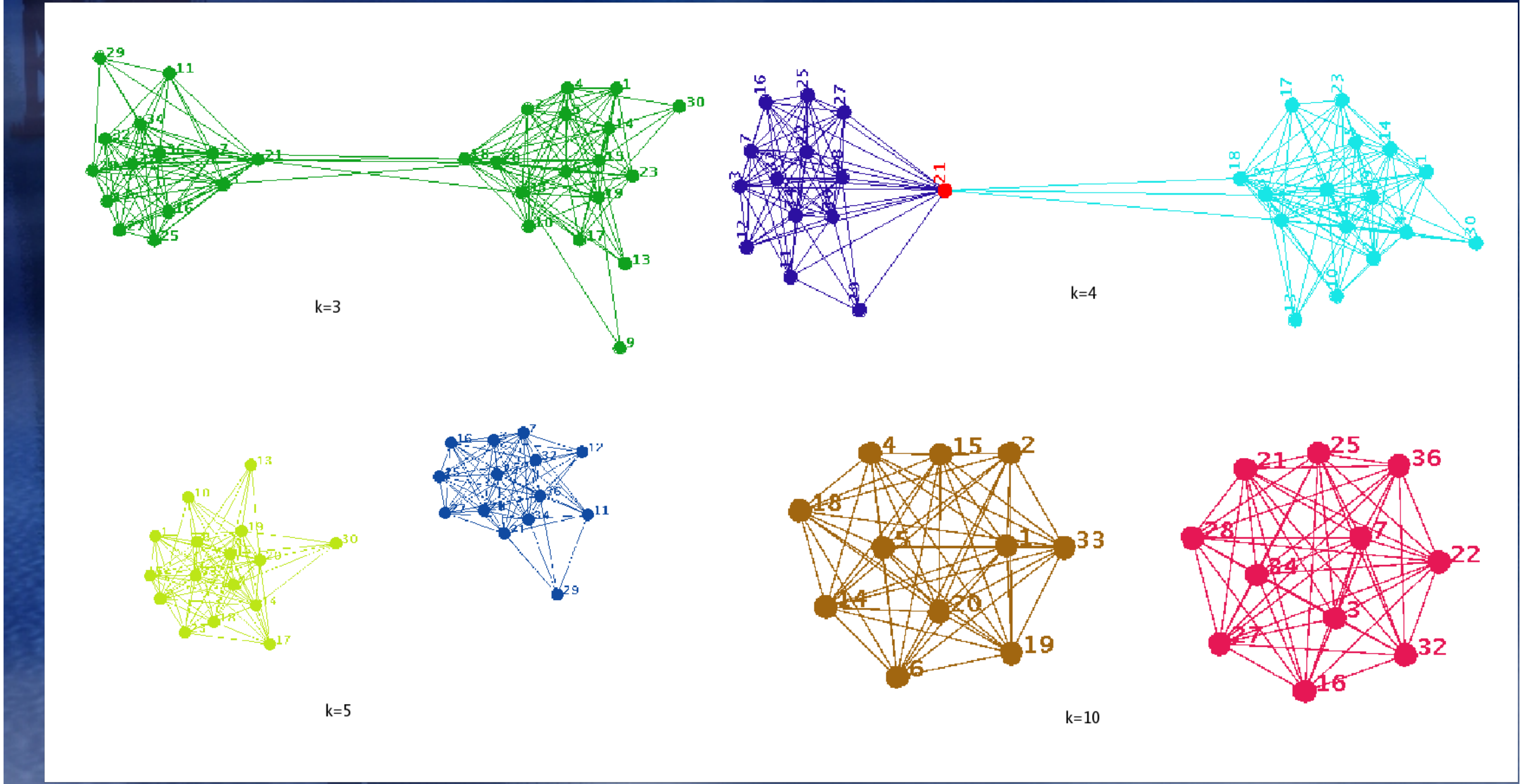
#name	port	Parrot	Wooden leg	Height	
xxx	penzance	y	y	1.75	
yyy	penzance	y	y	1.74	
zzz	penzance	y	y	1.76	
Dread pirate roberts (*)	?	n	n	1.80	
foo	neverland			1.65	
bar	airport			1.60	
baz	hollywood			1.50	
fie	Treasure island	y	y	2.00	

Adding the graph messes this all up

- Link data represents a lot of attacks on hash of name:



Worse: K-Clique Analysis


k=3

k=4

k=5

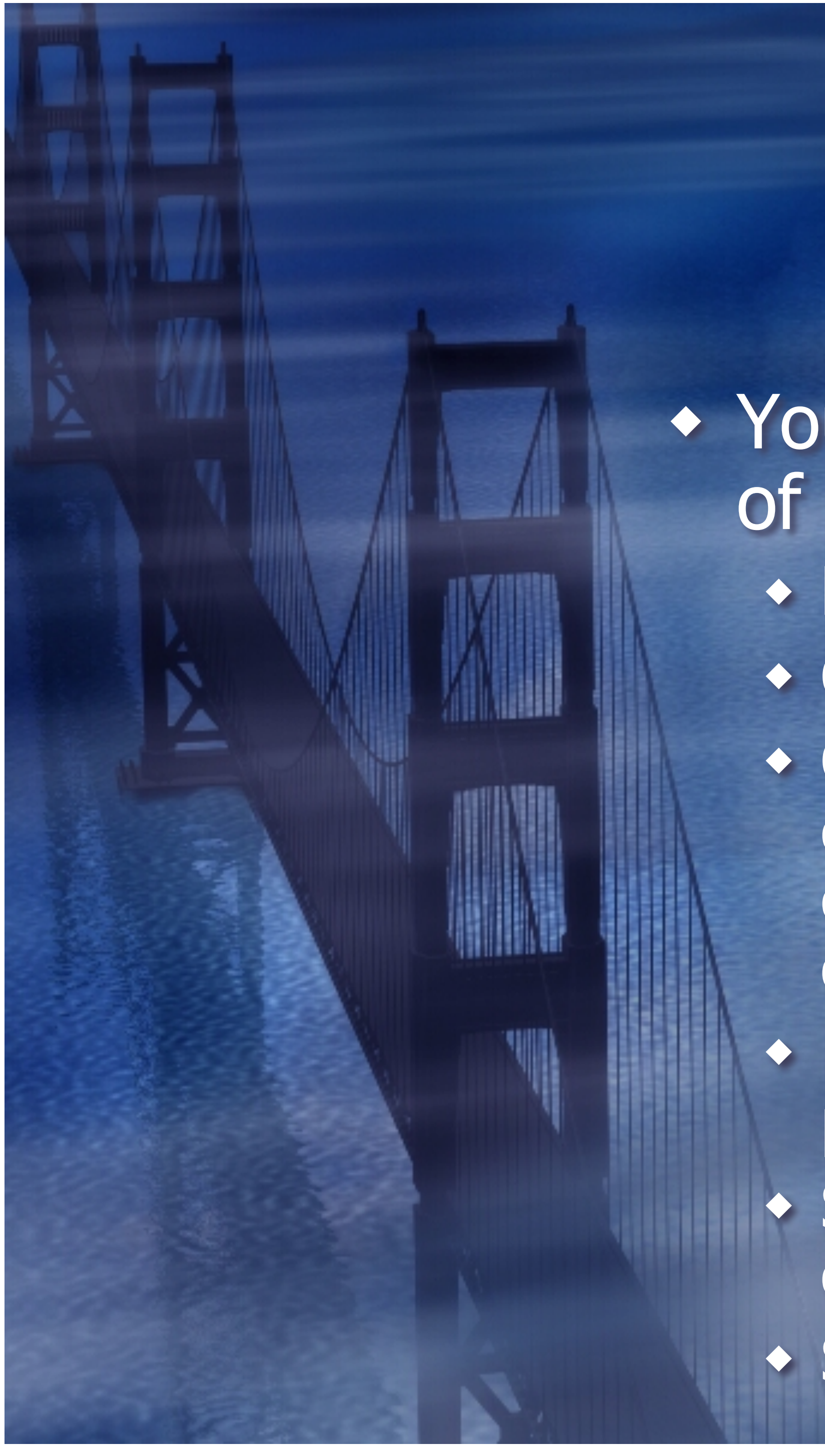
k=10

There are lots of 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"



Hypergraphs

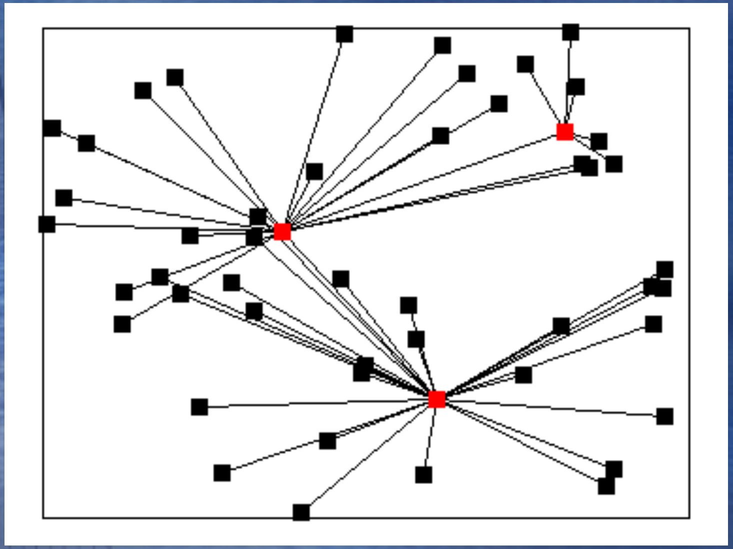
- You have an edge for each type of relationship
  - kin, friend, colleague
  - Co-author of work
  - Co-located (e.g. paid congestion charge same time, used oyster card on same journey, checked in on foursquare same place)
  - Pay tax together, live at same postcode,
  - Sent SMS, IM, Email, Phone call,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

Re-identification 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…..

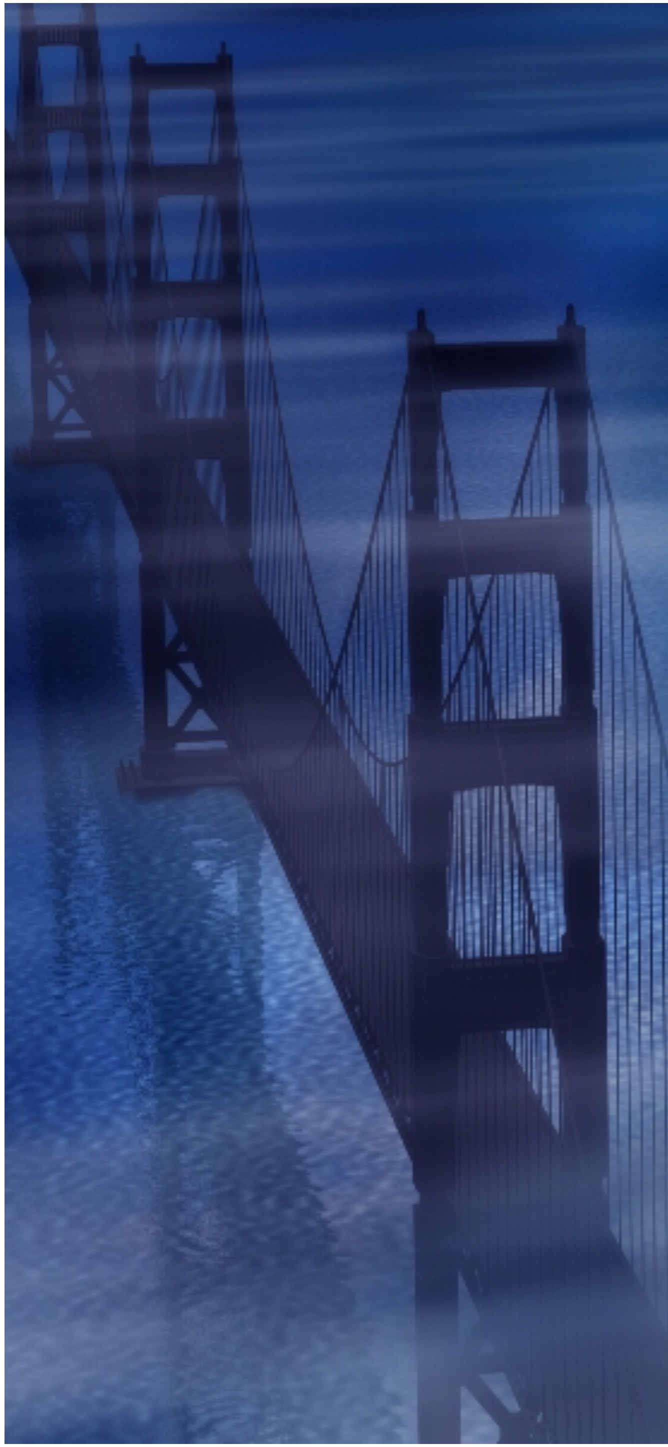
Dynamics

- Forgetting might help



Manifesto

- Separate storage of node PII and link data
- Always crypt PII
- Decentralize nodes *and* links
- Partition PII by role
  - Kin, friend, worl, school
  - Health, finance, gov, social
- Make it easy to understand
  - Maybe add forgetting



Take Homes

- Doesn’t have to be all central
  - Cannot figure out safe way to share graphs (sorry:-(
  - Can use Differential Privacy for node data records (without graph)
  - Epidemiologists don’t need our bank data, government don’t need our social data
- Prototype by some colleagues at Eurecom☺

  http://www.safebook.us/