Privacy Leak in Egonets

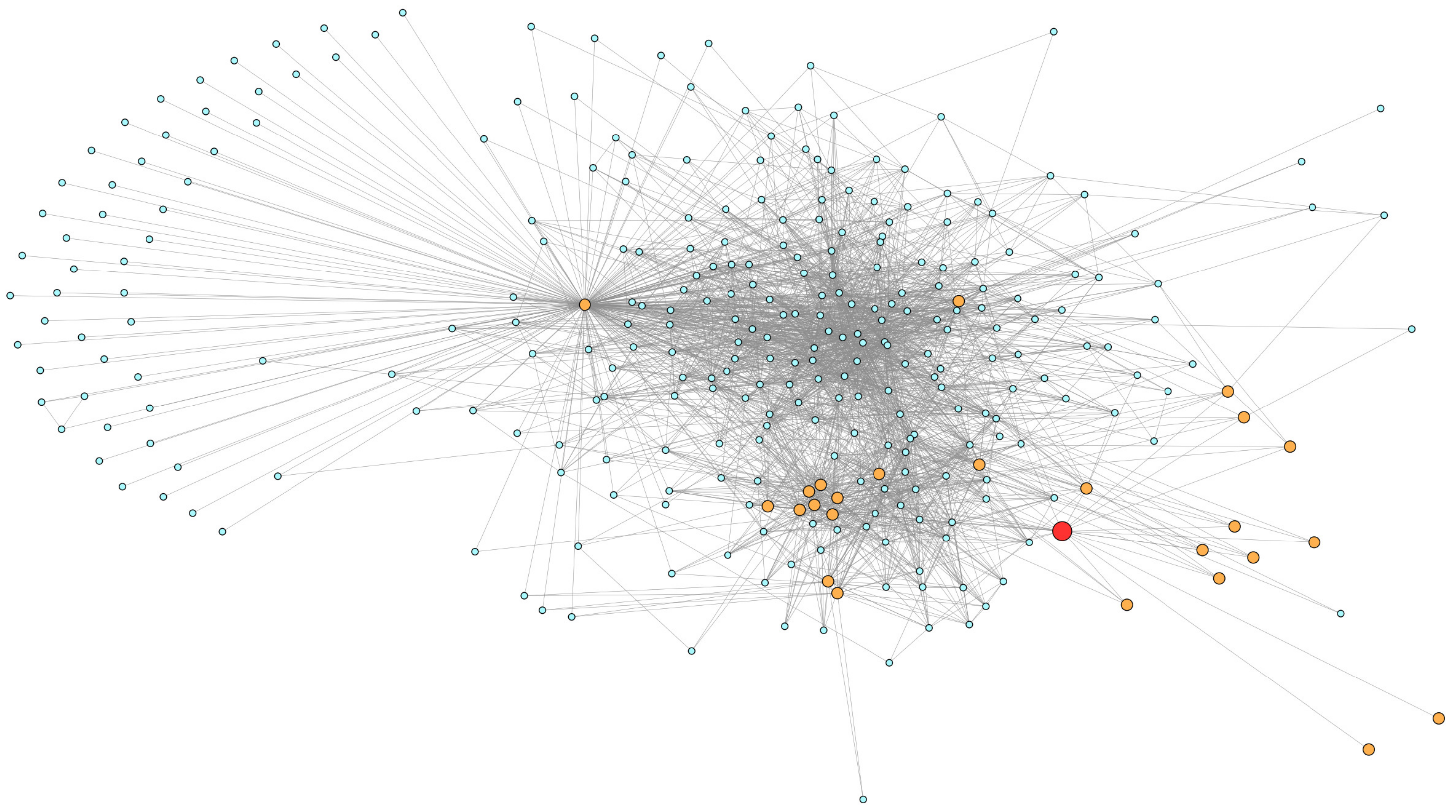
Kumar Sharad, Computer Lab, University of Cambridge
George Danezis, Microsoft Research Cambridge

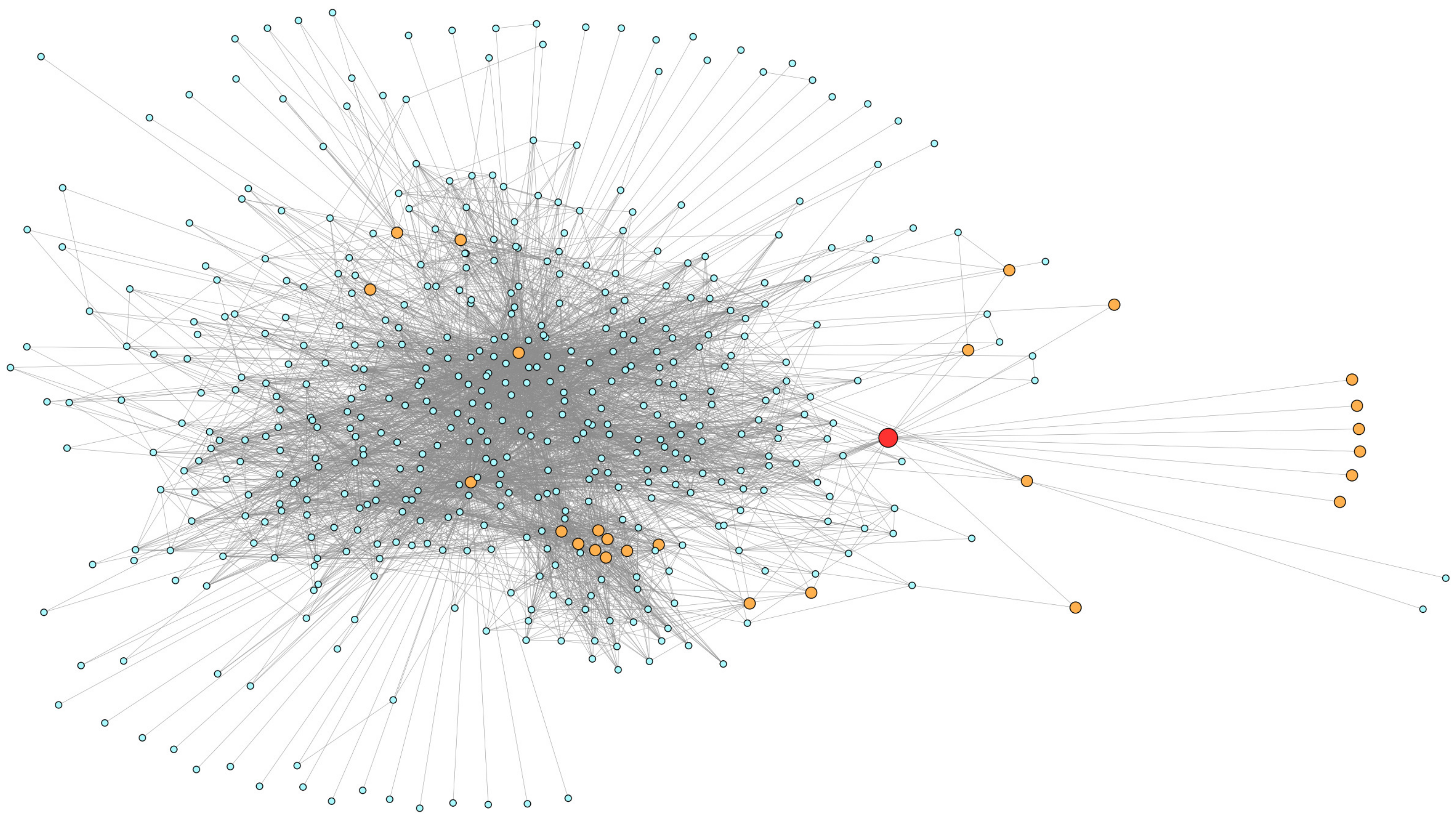
Abstract

A major Telecom released anonymized communication sub-graphs of a few thousand randomly selected subscribers for scientific analysis. To obfuscate the interactions between the sub-graphs the common nodes were renamed. However, this is not enough to preserve privacy.

Extracted Sub-graphs

The sub-graphs tend to have nodes in common which can be re-identified by analysing the graph topology.

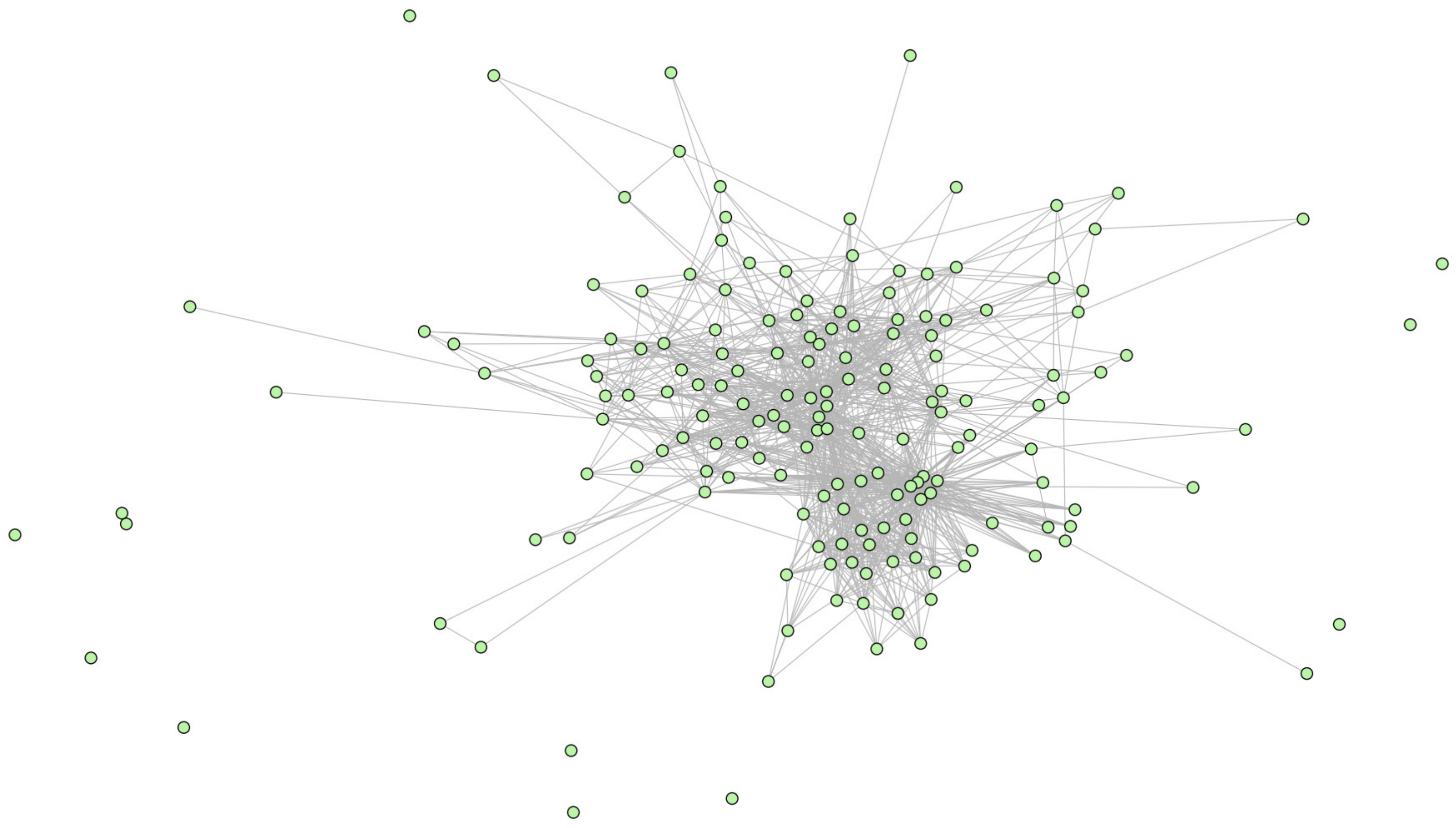




Recovering the full communication graph could lead to severe privacy breaches.

Re-identification

By observing the neighbourhood of nodes we can assign a match probability to pairs of nodes across ego-nets. The nodes at a distance of 1-hop from the ego are easier to re-identify.



After we have a match score between all possible pairs across two ego-nets, an optimal matching can be found.

Results

- Almost all the common 1-hop nodes were re-identified with over 98% success probability.
- A significant portion of the 2-hop nodes were re-identified with over 75% (often over 90%) success probability.

Conclusion

- It is hard to release high dimensional data whilst preserving privacy.
- Anonymizing real world data does not work in general.

http://www.cl.cam.ac.uk/~ks640/ Kumar.Sharad@cl.cam.ac.uk