De-anonymizing D4D Datasets

Kumar Sharad [1] George Danezis [2]

[1]University of Cambridge

[2]Microsoft Research

July 12, 2013

6th Workshop on Hot Topics in Privacy Enhancing Technologies 2013, Bloomington, Indiana, USA

Can Personally Identifiable Information be Anonymized?

- Research indicates that anonymyzing feature rich data is hard.
- In general it is not possible while preserving the usefulness of data.
- Release of real data presents an interesting opportunity to test the science.
- Encourages responsible data release.

Overview

The Data for Development (D4D) Challenge[1]

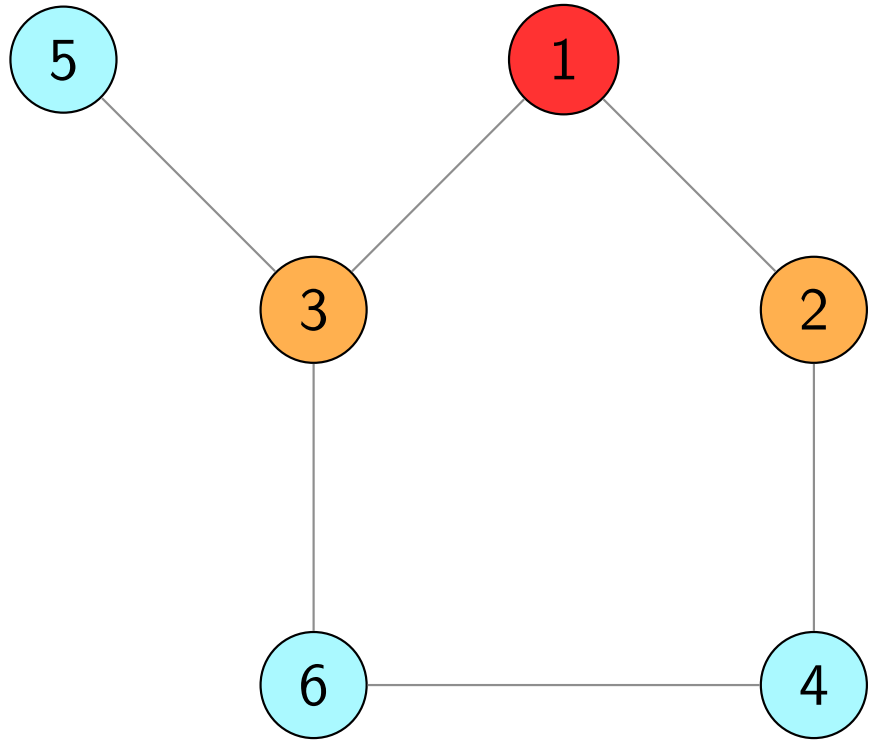
- Introduced by Orange in July 2012 for research related to social development in Ivory Coast.
- Four datasets of *anonymized* call patterns released.
- We were provided a preliminary version of the datasets.
- Ivory Coast facts
  - Population - 22.4 million.
  - Mobile phone users - 17.3 million.
  - Orange subscribers - 5 million.
  - A country fraught with civil war.

[1] http://www.d4d.orange.com/

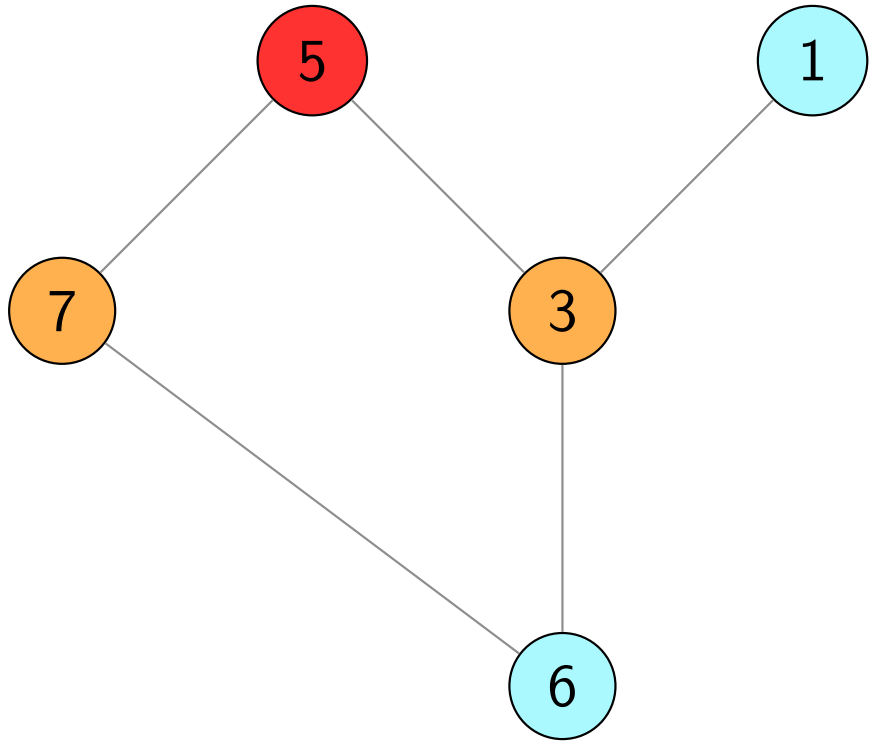
The Dataset 4

- Contains communication sub-graphs (ego nets) of 8300 randomly selected individuals (egos).
- Provides all communication between egos and their neighbours upto 2 degrees of separation.
- All nodes have random identifiers.
- Nodes common between sub-graphs have a different identifier in each sub-graph.

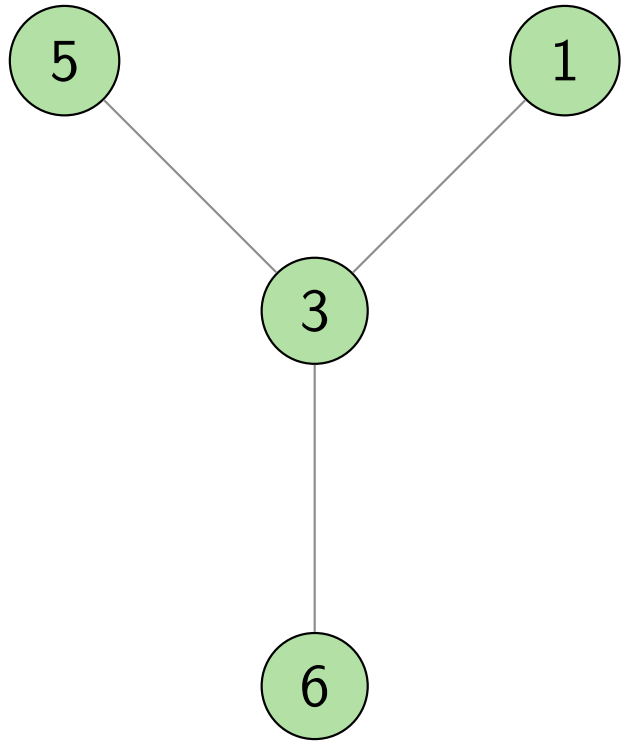
Toy Example



The ego net G_0

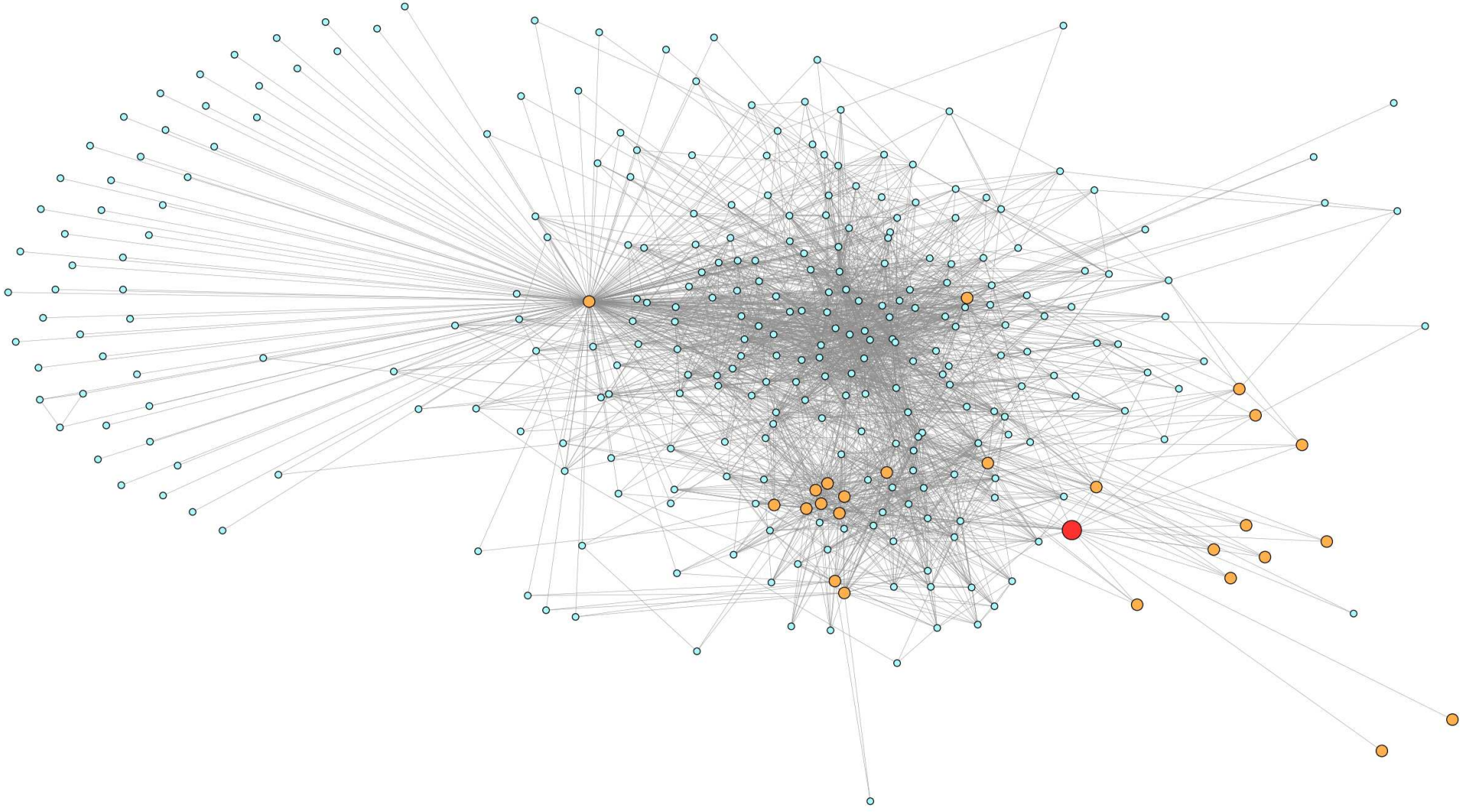


The ego net G_1

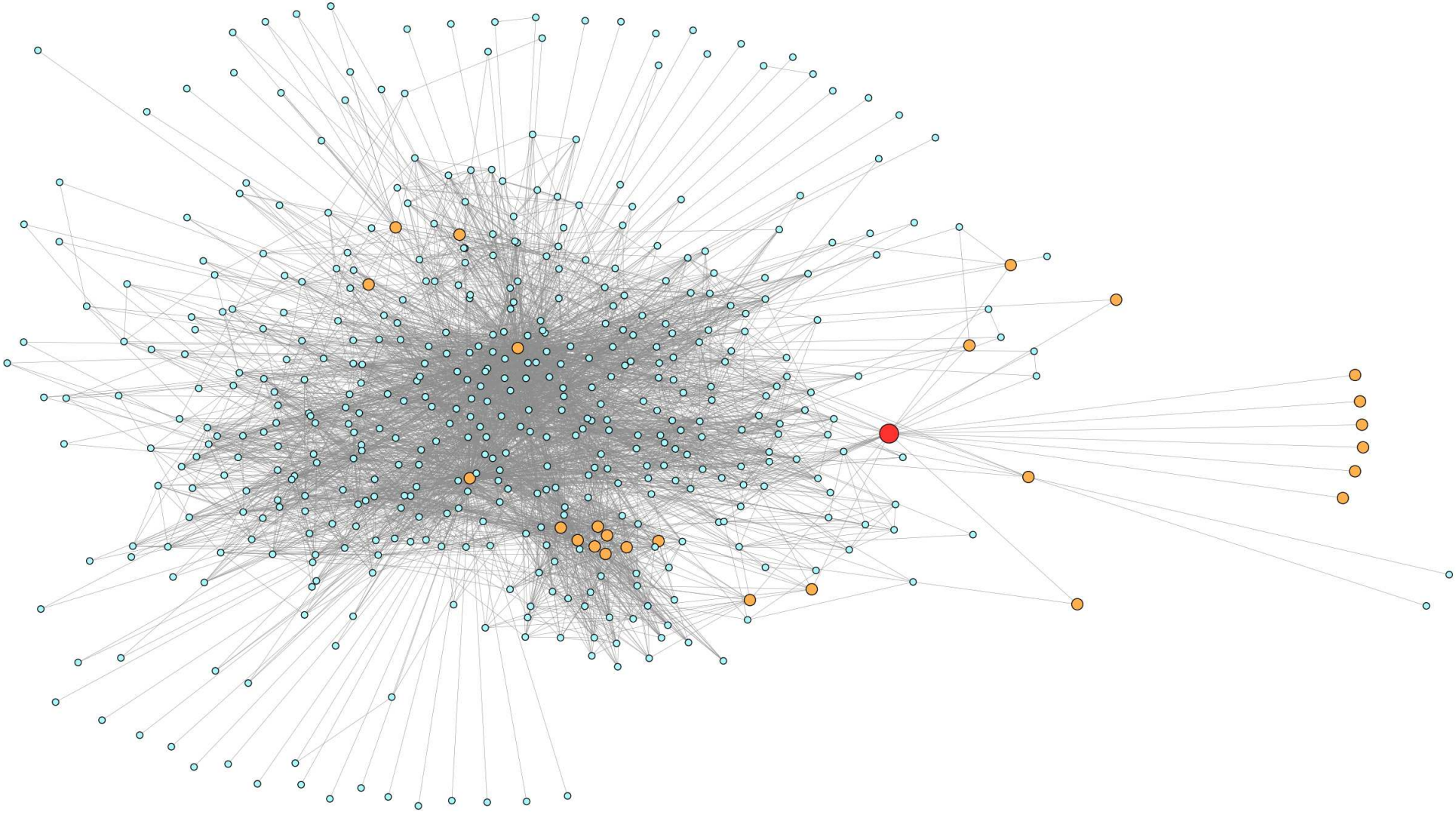


Sub-graph common to both G_0 and G_1

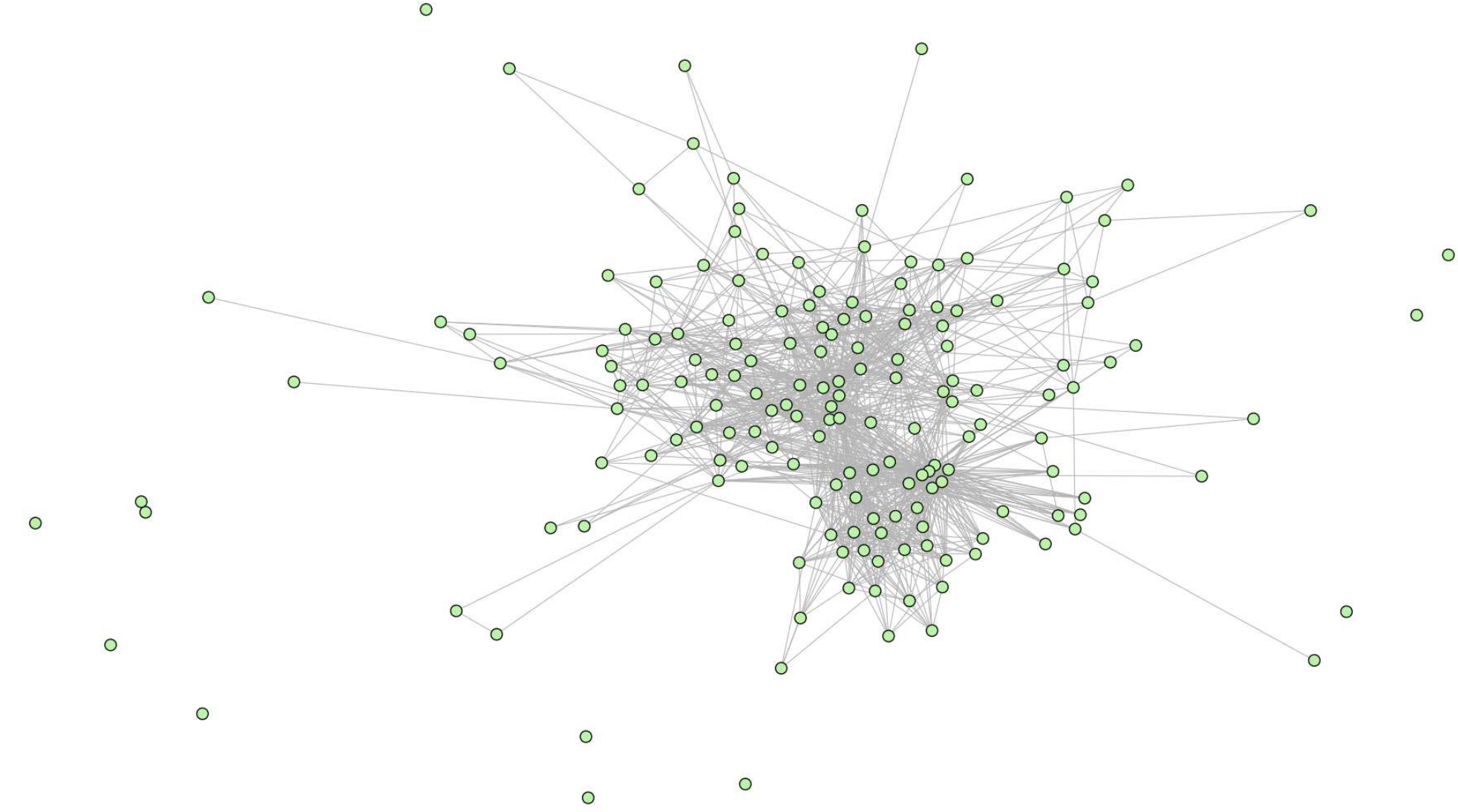
Real World Example



The ego net G_0



The ego net G_1



Sub-graph common to both G_0 and G_1

Re-identification

1-hop nodes

- Complete neighbourhood graph available.
- The degree distribution of a node's neighbours is almost unique.
- Graph invariants *completely* preserved even after anonymization!
- Use this to map nodes across ego nets.

2-hop nodes

- Parts of neighbourhood graph missing.
- Graph invariants *partially* preserved after anonymization.
- Observe the 1-hop nodes common between a pair of nodes in two ego nets.
- For pairs with significant match, find the cosine similarity between them based on the degree distribution of neighbourhood.
- Use bipartite matching to maximize the overall similarity score across pairs.

Results[2]

1-hop nodes

- Almost all the common nodes were re-identified with over 98% success rate.
- Hard to identify secluded nodes.

2-hop nodes

- Close to 15% (often over 20%) of common nodes re-identified.
- Success rate over 75% (occasionally over 90%).

[2] Based on EU email communication network - http://snap.stanford.edu/data/email-EuAll.html

Open Problem

How to efficiently re-identify nodes across ego nets which have no 1-hop nodes in common?

Contact

Kumar Sharad

- Kumar.Sharad@cl.cam.ac.uk
- research.sharad.de

George Danezis

- gdane@microsoft.com
- research.microsoft.com/en-us/um/people/gdane