Graph Anonymization

The impact of random perturbation in social networks

Francesco Stucci

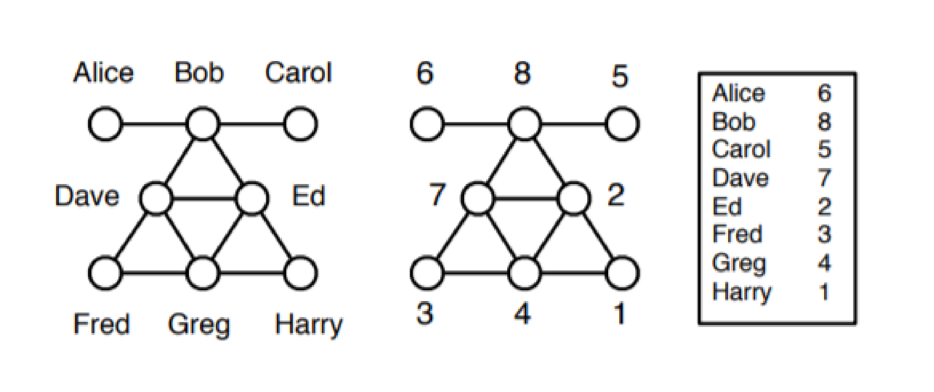
What «Anonymous» means:

lacking individuality, distinction, or recognizability

In graphs, in particular social networks, it means the incapability to associate each node to a person.
In this scenario we describe people only with labels (their names) by treating them as our sensitive information. We are not going to consider quasi-identifier information.

Naive Anonymization

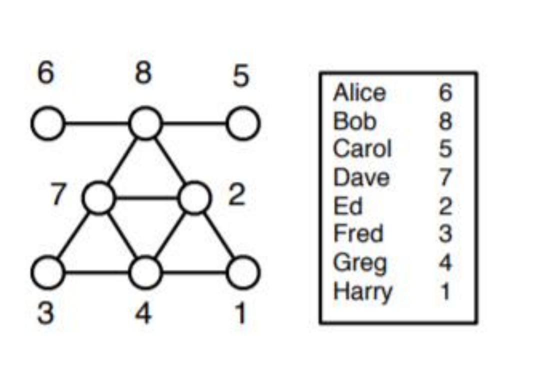
The most trivial way to anonymize graphs is by replacing identifiers with numbers



What about external information?

By knowing some information about nodes, adversaries coul be able to identify people from graph

Let's suppose we know our target has 1 friend (1 edge).
From such a graph we can conclude our target is in {5, 6} which has a very high probability of re-identification



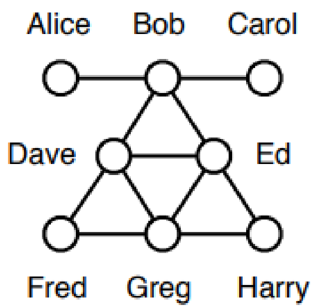
Adversary Knowledge

We define two classes of knowledge queries available to an adversary

- Vertex refinement queries
- Subgraph knowledge queries

Vertex Refinement

$$H_i(x) = \{H_{i-1}(n_1), H_{i-1}(n_2). . . , H_{i-1}(n_m)\}$$



(a) graph

Node ID	$\mathcal{H}_0$	$\mathcal{H}_1$	$\mathcal{H}_2$
Alice	ϵ	1	$\{4\}$
Bob	ϵ	4	$\{1, 1, 4, 4\}$
Carol	ϵ	1	$\{4\}$
Dave	ϵ	4	$\{2, 4, 4, 4\}$
Ed	ϵ	4	$\{2, 4, 4, 4\}$
Fred	ϵ	2	$\{4, 4\}$
Greg	ϵ	4	$\{2, 2, 4, 4\}$
Harry	ϵ	2	$\{4, 4\}$

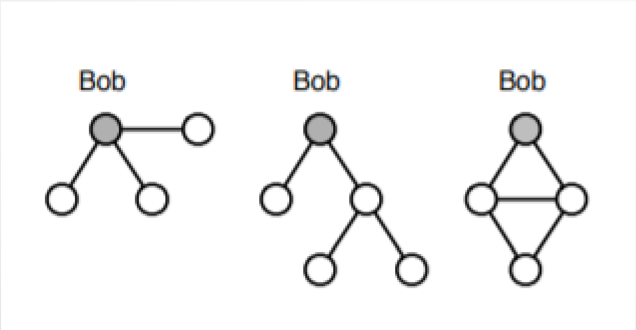
(b) vertex refinements

Equivalence Relation	Equivalence Classes
$\equiv_{\mathcal{H}_0}$	$\{A, B, C, D, E, F, G, H\}$
$\equiv_{\mathcal{H}_1}$	$\{A, C\}$ $\{B, D, E, G\}$ $\{F, H\}$
$\equiv_{\mathcal{H}_2}$	$\{A, C\}\{B\}\{D, E\}\{G\}\{F, H\}$
$\equiv_{A}$	$\{A, C\}\{B\}\{D, E\}\{G\}\{F, H\}$

(c) equivalence classes

Subgraph Knowledge

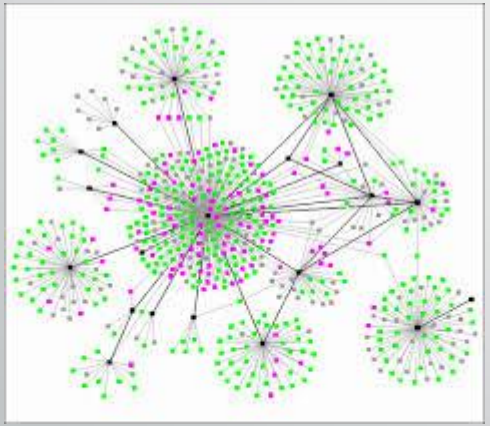
With subgraph knwoledge we define our queries by counting the edge in the subgraph. We re\fer to these as **Edge Factors**



Three instance of Bob node subgraphs with respectively 3, 4 and 4 edge factor

Used Graph

For this experiment we are going to use a Scale-Free network graph, which is a network whose degree distribution follows a power law.



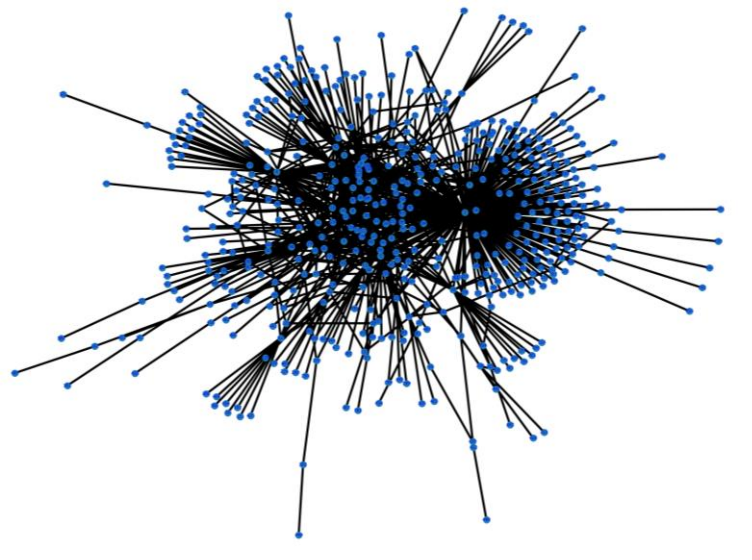
It's been chosen this kind of graph because its structure similarity with Social Network

Example graphs used in the paper are to big to deal with

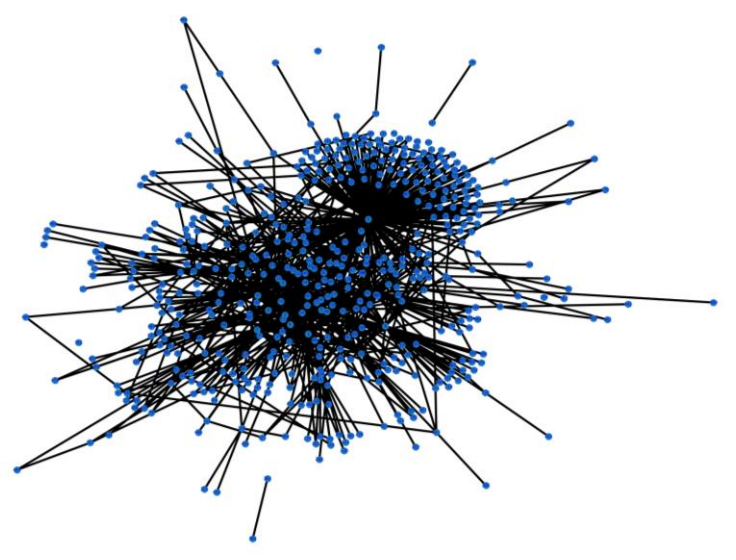
Let's make some tests

Just to have an idea on used graphs, this is the graph from which we have obtained our results.

We are going to test our de-anonymization technique on 0%, 0.2%, 0.5% and 10% perturbed graph

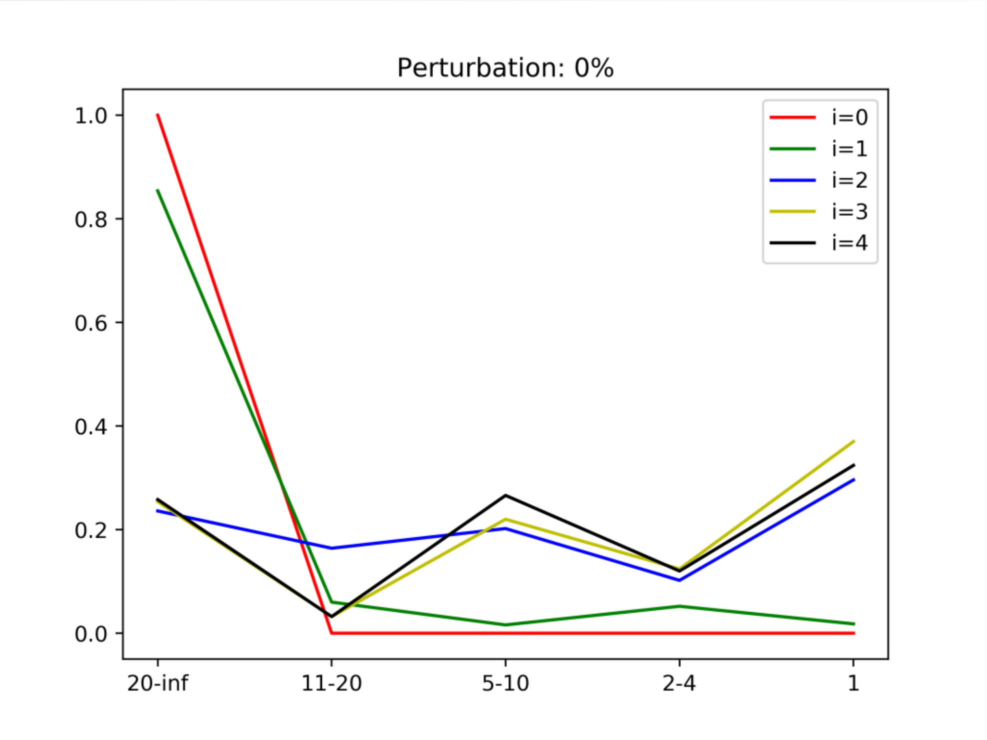


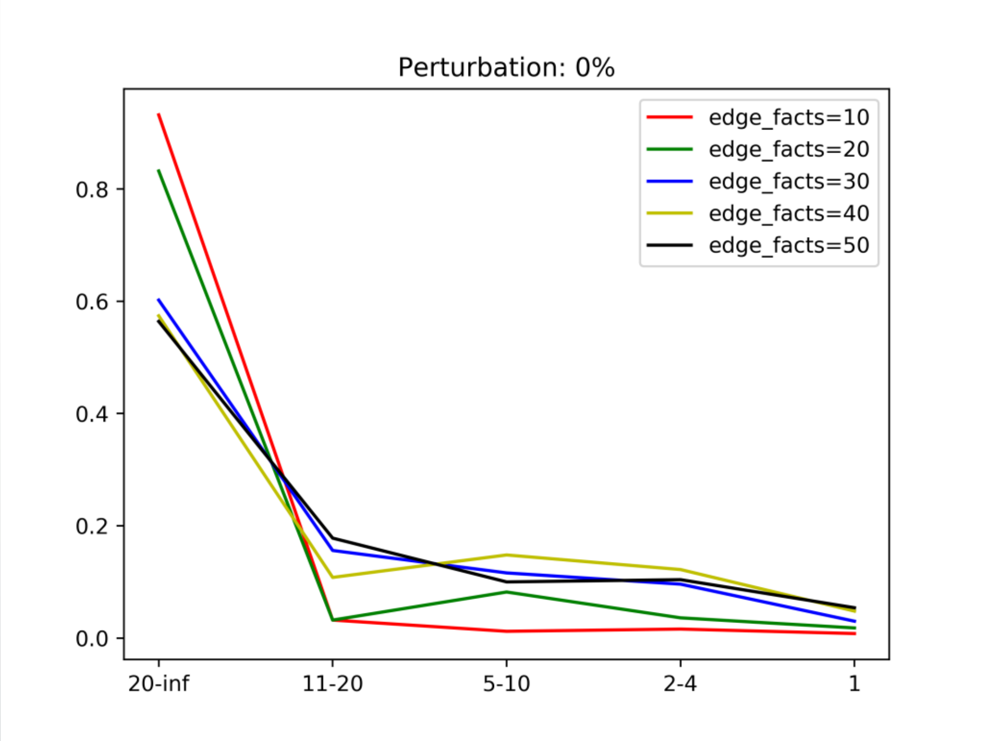
0% perturbation



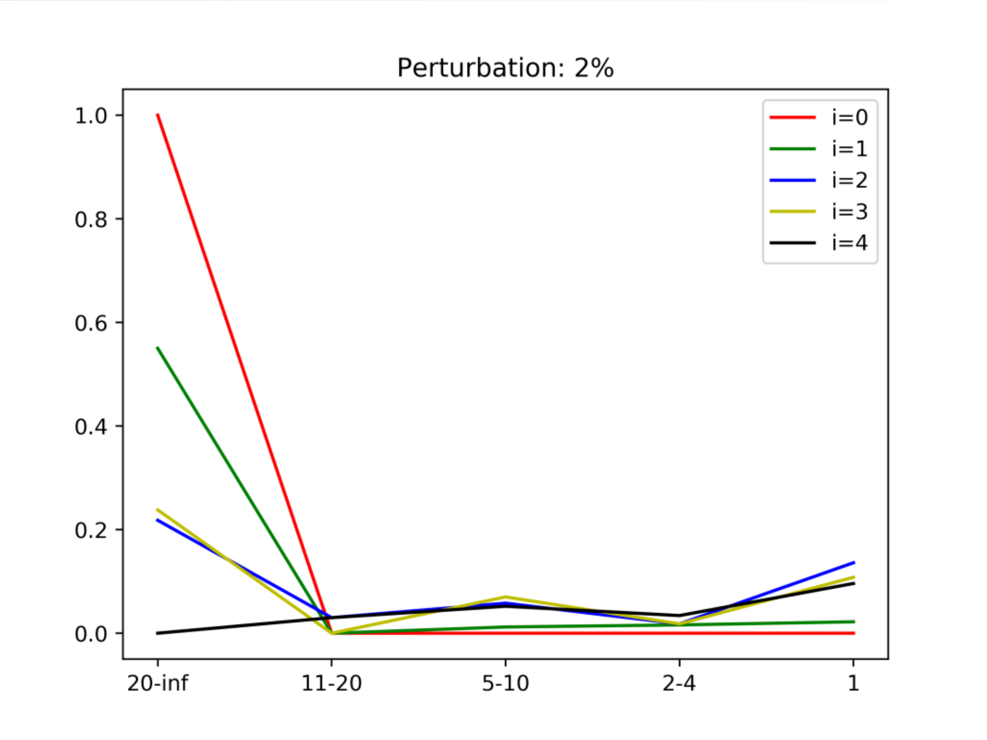
10% perturbation

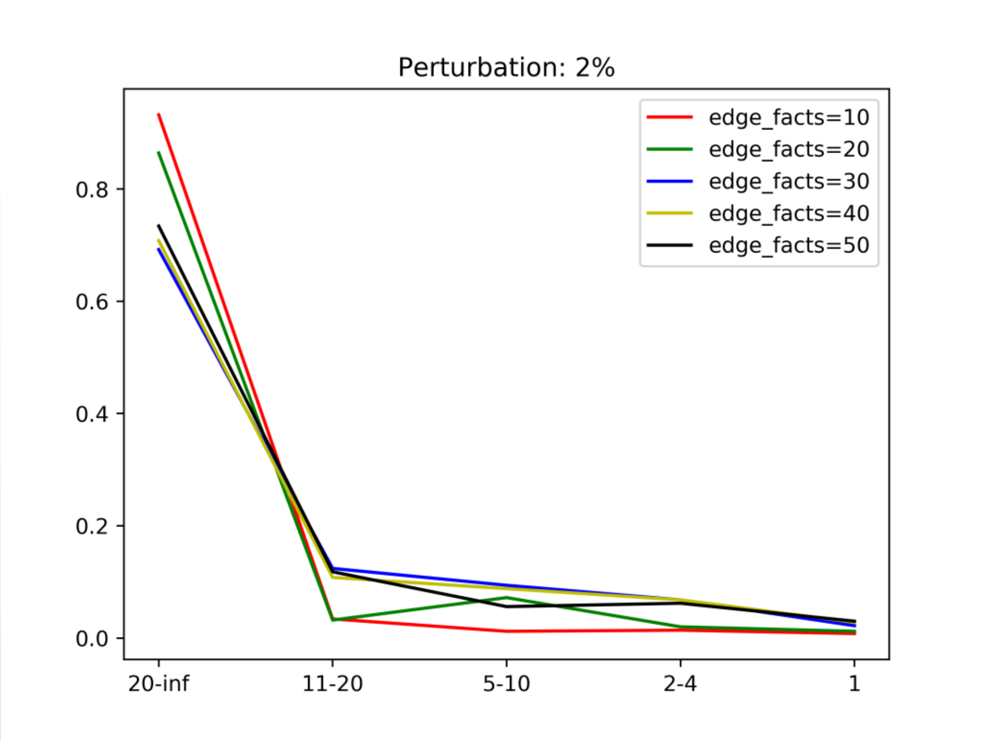
Results: 0% perturbation



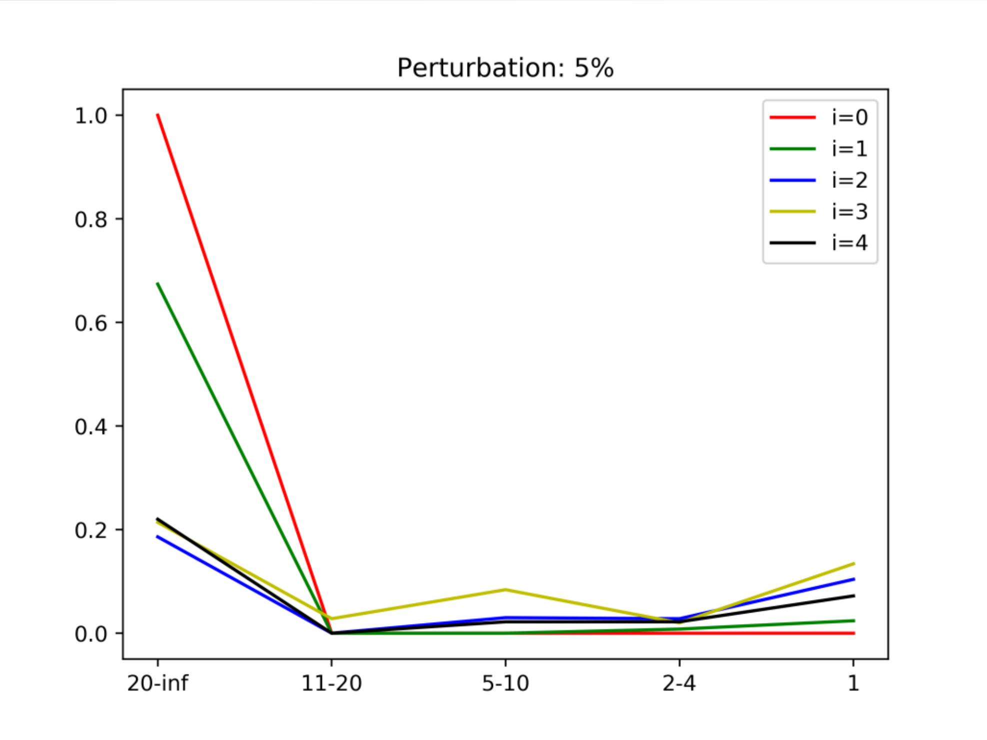


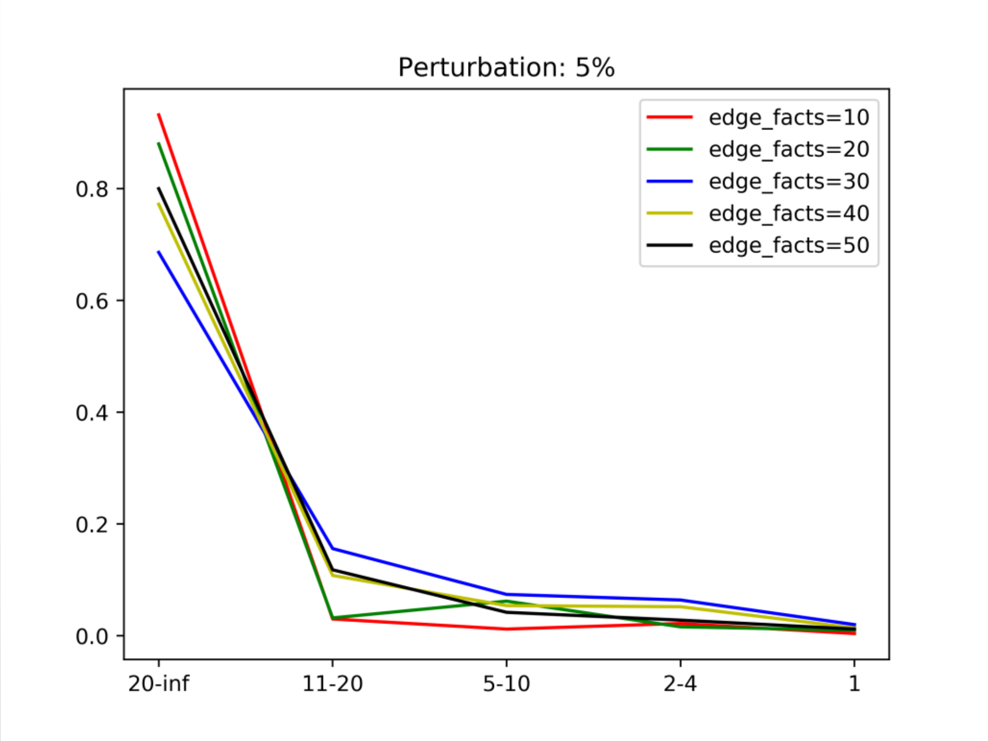
Results: 0.2% perturbation



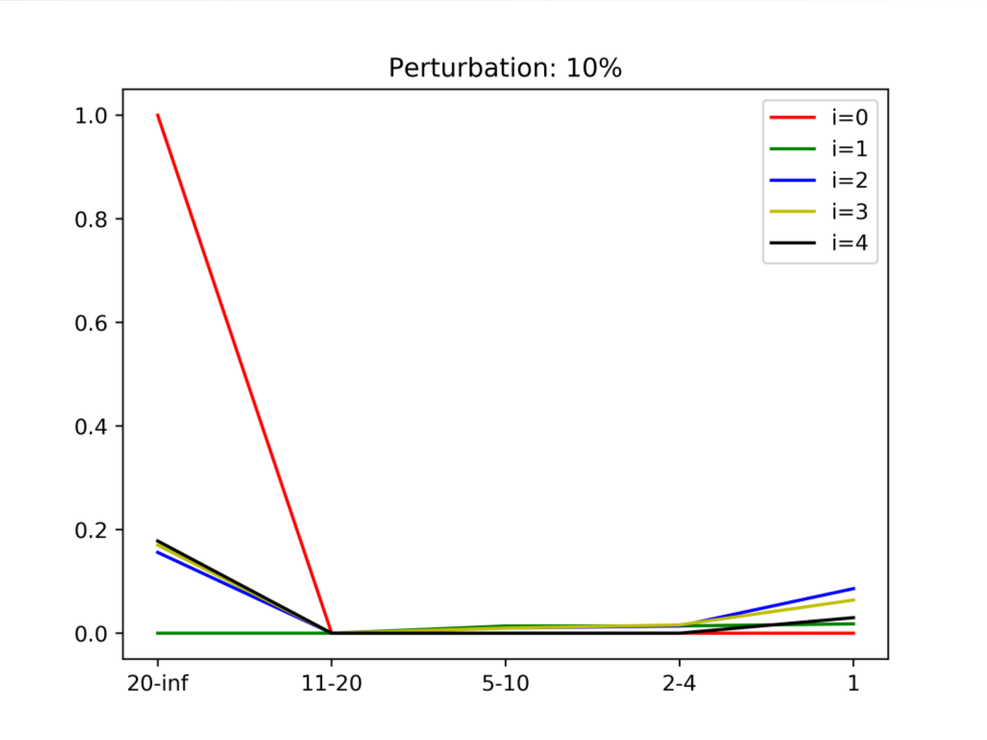


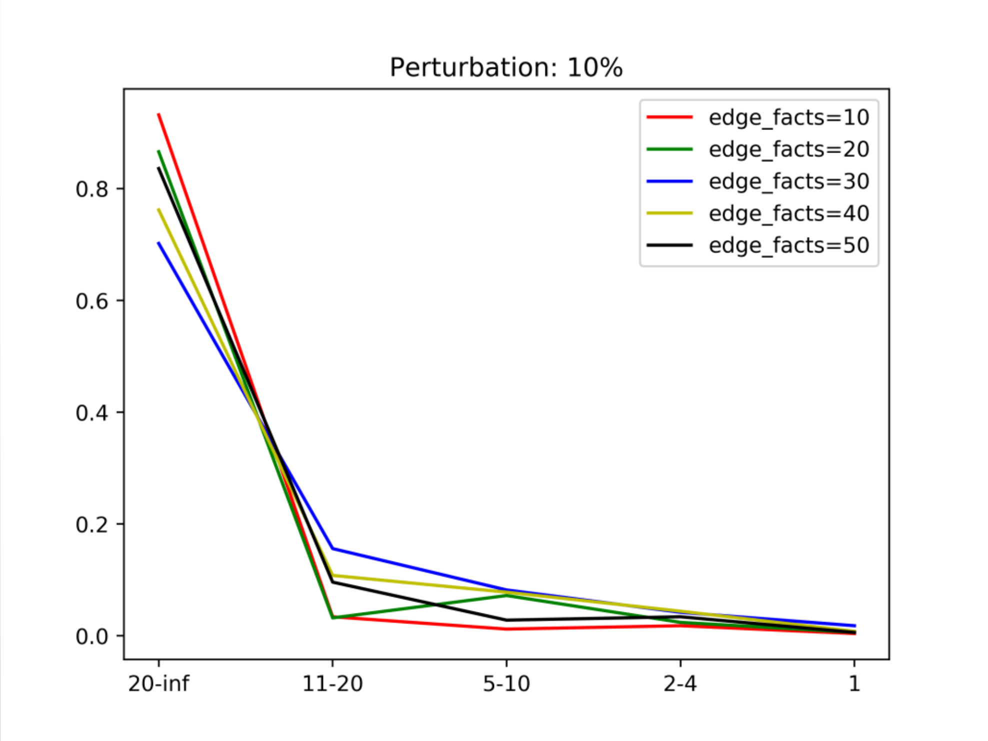
Results: 0.5% perturbation





Results: 10% perturbation





Some words about Information Loss

By pertrubating randomly a graph, we lose some information.

While the perturbed graphs are often distinct from a completely random graph, the information loss after a perturbation of 10% of the edges appears to be substantial

	Enron			
Measure	Original	Perturbed 5%	Perturbed 10%	Random (100%)
Degree	5.0	4.5	4.6	5.0
Diameter	9.0	8.7	7.6	6.1
Path length	4.0	3.2	3.0	3.0
Closeness	0.276	0.293	0.304	0.337
Betweenness	0.005	0.009	0.010	0.014
Clust. Coeff.	0.286	0.242	0.191	0.000

The Enron graph features changes based on perturbation

Model based perturbation

A strategy for maintaining accuracy under perturbation is for the data trustee to derive a statistical model of the original data, and to use that model to "bias" the random perturbation towards those that respect properties of the graph

Conclusions

- We showed the behaviour of two types of adversary knowledge query on a naive graph: without any kind of perturbation, a good portion of nodes can be de-anonymised
- We tried some percentage of pertrubation in order to minimize the number of de-anonymized nodes
- We found a good trade-off between anonymization and utility loss