UNIVERSITÄT SIEGEN

Security of Online Social Networks

Social Network Analysis II

Lehrstuhl IT-Sicherheitsmanagment

Universität Siegen

May 31, 2012

Overview Lesson 07

Clustering

- Clustering Quality
- k-means

Graph Matching

- De-Anonymizing Social Networks

Clustering

Overview Clustering

- Classification of Samples
- Profiling
- Finding “Groups”
- “Hidden” similarities
- e.g. people most likely inclined to buy X
- e.g. behavioral anomalies

Clustering Example

- Clustering: Partition Π of a set
- Partitioning of a birthday cake (without measure)
- e.g. determining "supporters" (Example)

Clustering Quality Measures

Quality Criteria

- External (absolute)
  - using information not available to algorithm
  - "Cake-Criteria": Equivalent Partition-Size
- Internal (relative)
  - allows for $Q(\Pi') < Q(\Pi'')$
  - e.g. silhouette coefficient

Silhouette Coefficient I

Internal Quality Criterion for Quality of Placement of Sample $s(a)$

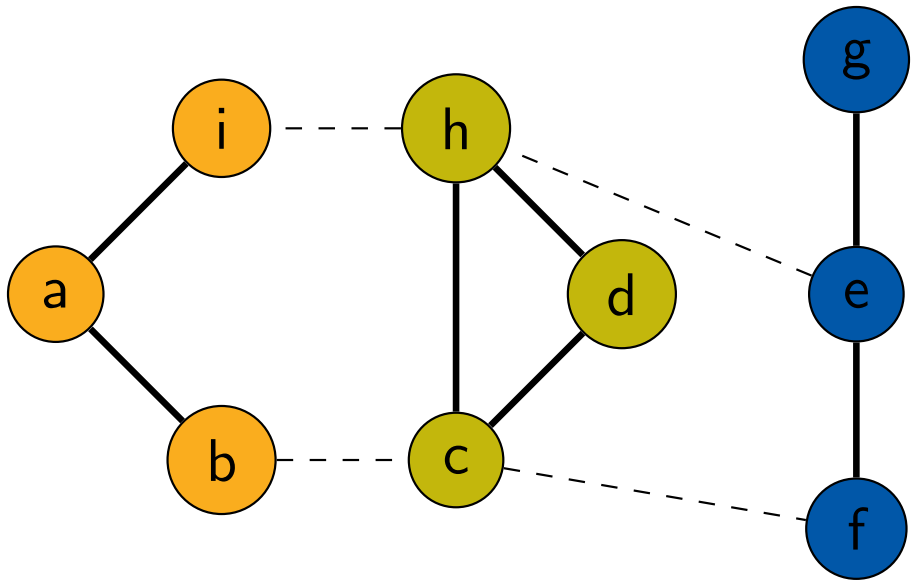
Clustering $\Pi = \{\pi_1, \ldots, \pi_k\}, a \in \pi \subseteq \mathcal{A}$

$$
\begin{aligned}
I(a) &:= \frac{1}{|\pi_i|} \sum x \in \pi_i d(x, a) \\
O_j(a) &:= \frac{1}{|\pi_j|} \sum x \in \pi_j d(x, a) \\
O(a) &:= \min\{O_1(a), \ldots, O_{i-1}(a), O_{i+1}(a), \ldots, O_k(a)\} \\
s(a) &:= \frac{O(a) - I(a)}{\max\{O(a), I(a)\}}
\end{aligned}
$$

Silhouette Coefficient II

- Partition: $\underset{a \in \pi}{mean}\{s(a)\}$
- Clustering: $\underset{\pi \in \Pi}{mean}\{s(\pi)\}$

Example:



	I	O	s
a	$2/2$	$min\{7/3, 10/3\}$	$4/3 * 3/7 = 4/7$
b	$3/2$	$min\{5/3, 9/3\}$	$1/6 * 2/3 = 1/9$
i	$3/2$	$min\{5/3, 8/3\}$	$1/6 * 2/3 = 1/9$
c	$2/2$	$min\{6/3, 6/3\}$	$1 * 2 = 2$
d	$2/2$	$min\{7/3, 7/3\}$	$4/3 * 3/7 = 4/7$
h	$2/2$	$min\{5/3, 5/3\}$	$2/3 * 3/5 = 2/5$
e	$2/2$	$min\{5/3, 8/3\}$	$2/3 * 3/5 = 2/5$
f	$3/2$	$min\{5/3, 8/3\}$	$1/6 * 2/3 = 1/9$
g	$3/2$	$min\{8/3, 11/3\}$	$7/6 * 3/8 = 7/16$

	π_1	π_2	π_3
s	$50/189 \approx 0.26455$	$104/105 \approx 0.9905$	$683/2160 \approx 0.3162$

$s(\Pi) \approx 0.523751$

k-means

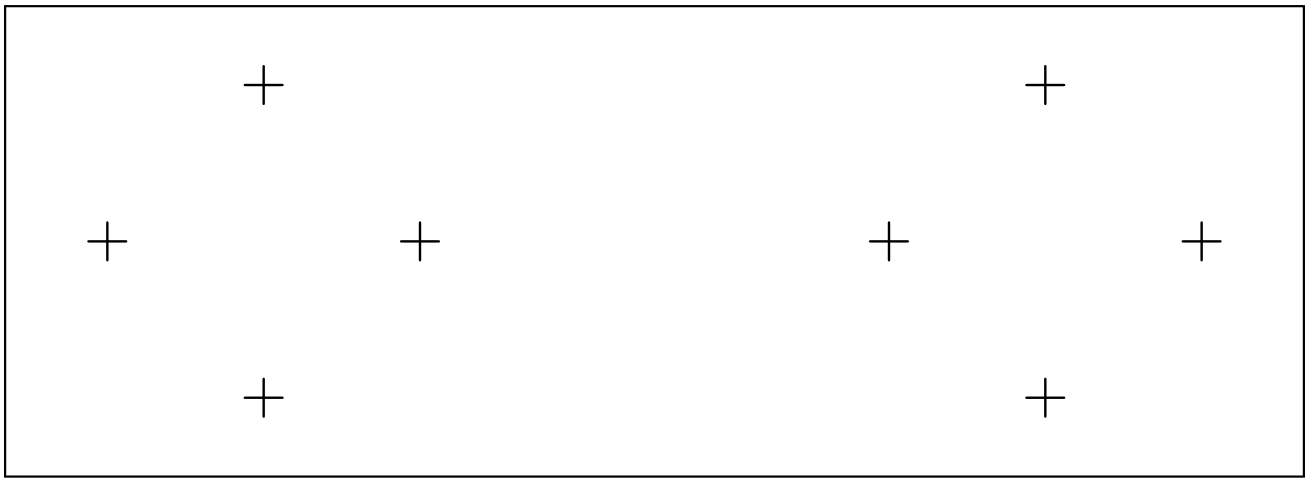
Basic k-means:

- samples $\mathcal{A}$
- k centroids $c_1, \ldots, c_k \in \mathcal{C}$

1. elements closest to a centroid belong to that partition
   - $\pi_i = \{a : a \in \mathcal{A}, d(c_i, a) \leq d(c_l, a), \text{ for } l \in \{1, \ldots, k\}\}$
2. adapt centroids to partition
   - $c(\pi_i) = \arg\min_{x \in \mathcal{C}} \left\{ \sum_{a \in \pi_i} d(x, a) \right\}$
3. iterate until sufficient quality

Example k-means

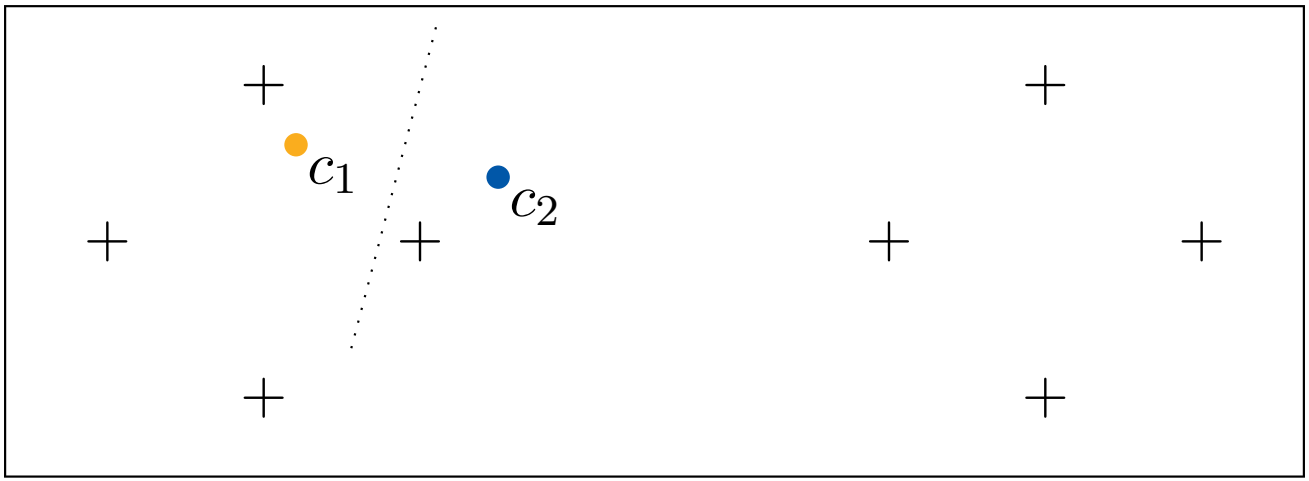
$k = 2$



$$\pi_i = \{a : a \in \mathcal{A}, d(c_i, a) \leq d(c_l, a), \text{ for } l \in \{1, \ldots, k\}\}$$

Example k-means

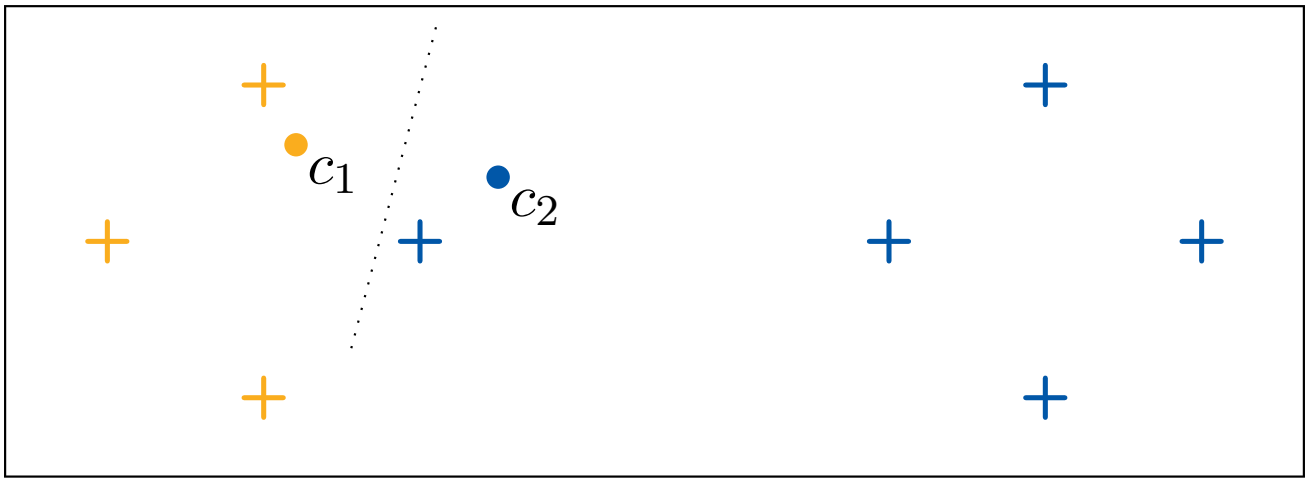
$k = 2$



$$\pi_i = \{a : a \in \mathcal{A}, d(c_i, a) \leq d(c_l, a), \text{ for } l \in \{1, \ldots, k\}\}$$

Example k-means

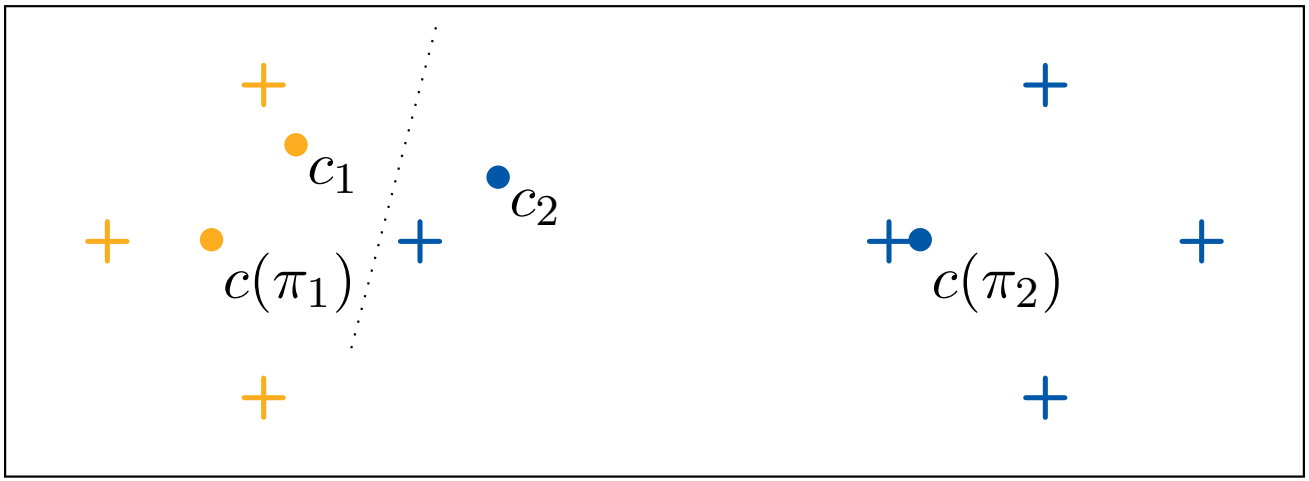
$k = 2$



$$\pi_i = \{a : a \in \mathcal{A}, d(c_i, a) \leq d(c_l, a), \text{ for } l \in \{1, \ldots, k\}\}$$

Example k-means

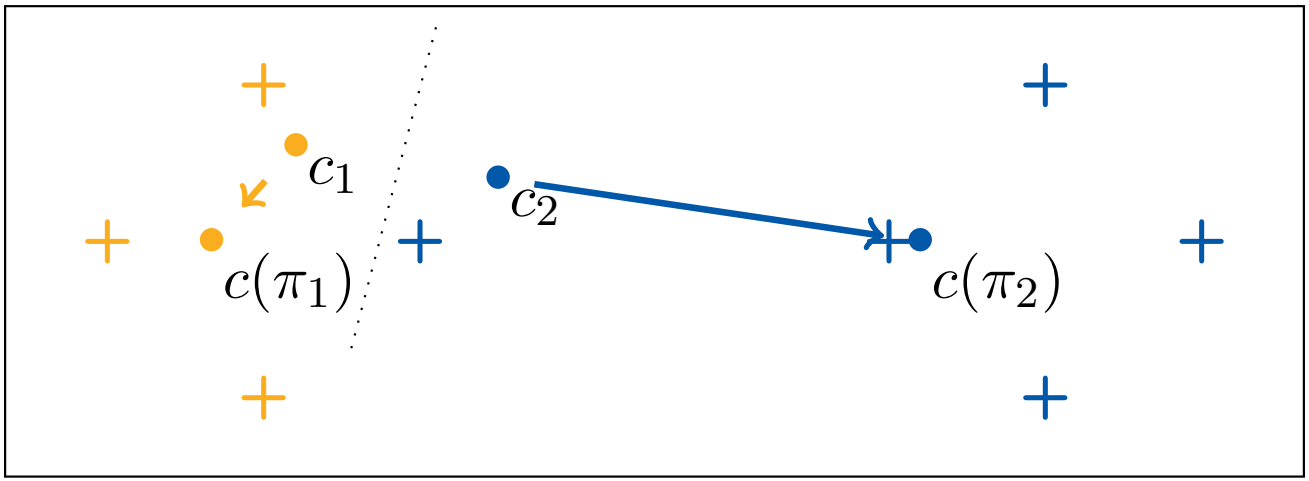
$k = 2$



$$c(\pi_i) = \arg\min_{x \in \mathcal{C}} \{ \sum_{a \in \pi_i} d(x, a) \}$$

Example k-means

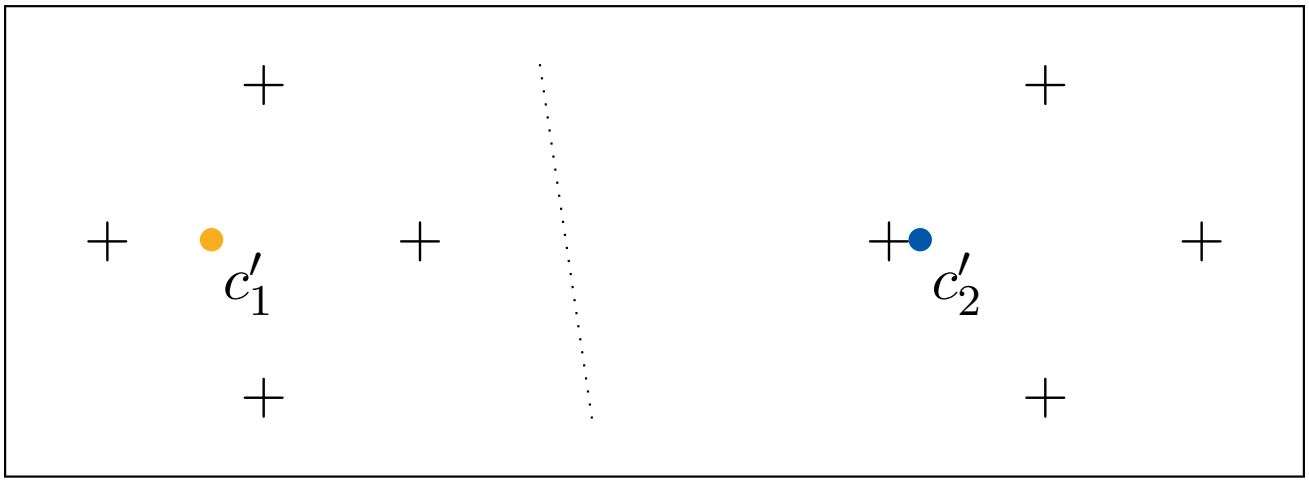
$k = 2$



$$c(\pi_i) = \arg\min_{x \in \mathcal{C}} \{ \sum_{a \in \pi_i} d(x, a) \}$$

Example k-means

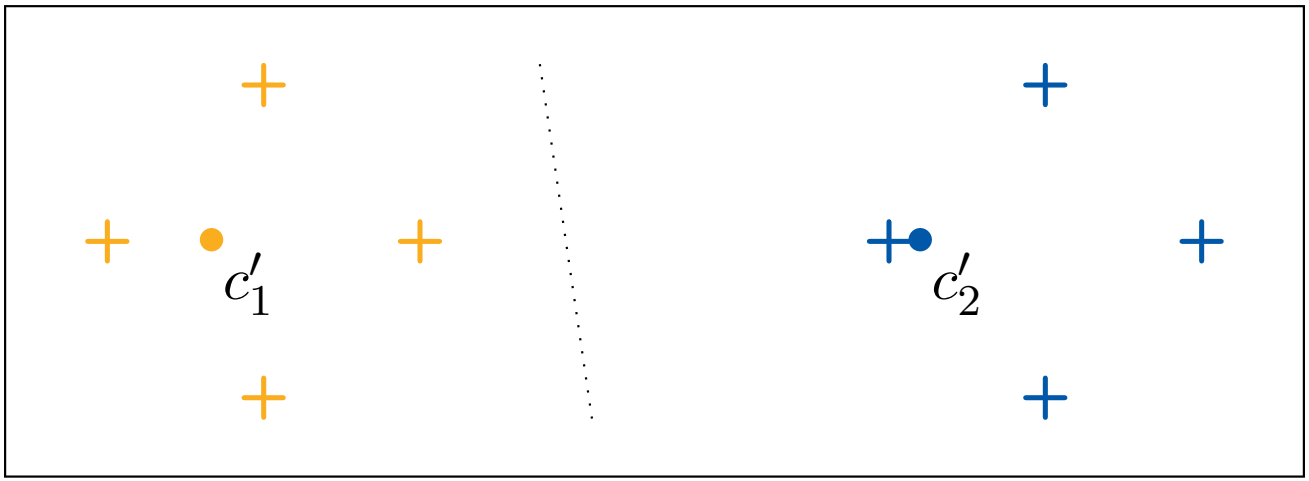
$k = 2$



Finish?

Example k-means

$k = 2$



Finish?

Example k-means

$k = 2$

Finish?

Example k-means

$k = 2$

Finish?

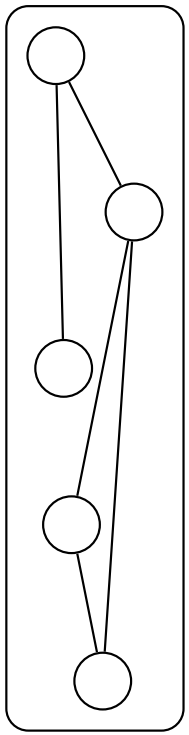
Clustering Algorithms

- Halting Problem
- Sensitive to initial conditions
- Local Maximum
- Batch vs. Incremental
- Large Collection of Algorithms:
  - k-means (Centroid-based)
  - Density-based
  - Distribution-based
  - Neuronal Networks
  - Hierachical Clustering (e.g. greedy least distance)
  - ...

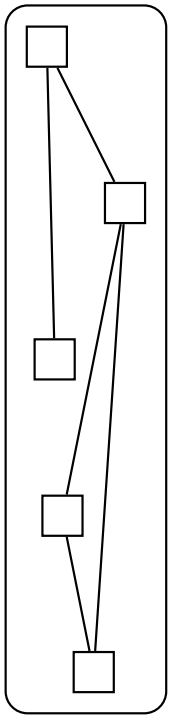
Graph Matching

Graph Matching

SocNet1

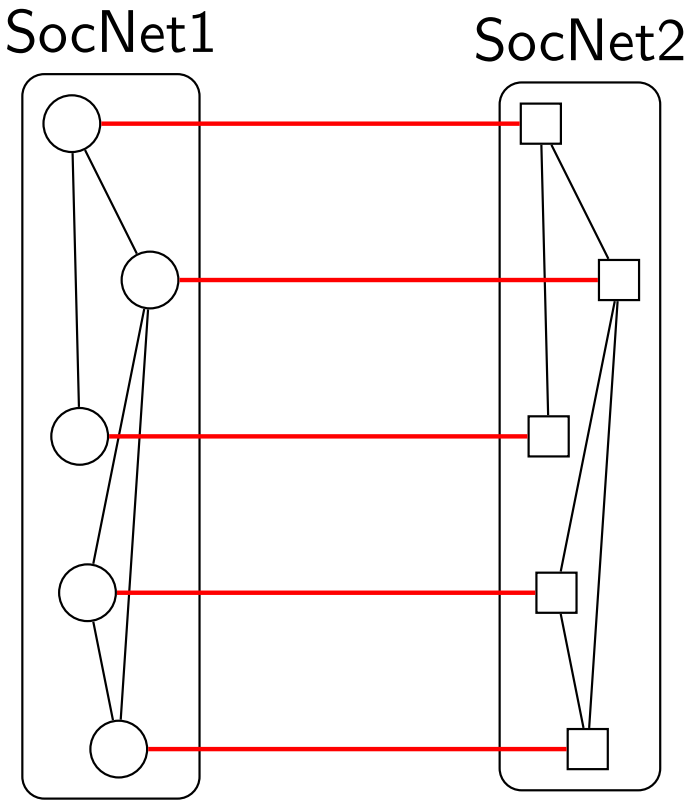


SocNet2



- Re-identification
- Imagine this with twitter and facebook
- Hypothesis Space Size? $n!$
- Improve your chances with seeds
- Iterate Nodes (n)

Graph Matching



- Re-identification
- Imagine this with twitter and facebook
- Hypothesis Space Size? $n!$
- Improve your chances with seeds
- Iterate Nodes (n)

De-Anonymizing Social Networks (2009)

- Narayanan, Shmatikov
- Sanitised $\mathcal{S}_{SAN}$: published, anonymised graph
- Auxiliary $\mathcal{S}_{AUX}$: different attributed graph
- Objective: Match sanitised onto auxiliary network
- Proceedure:
  1. Seed identification
  2. Propagation

Model & Definitions

- Social Network $\mathcal{S}$:
  - $G = (V, E, \mathcal{X}, \mathcal{Y})$: Nodes, Edges, Attributes
- Sanitisation/Data Release
  - Removing Attributes, Nodes, Edges
  - Fake Edges, ...
  - Sanitised graph $\mathcal{S}_{SAN}$
- Attack Scenarios
  - Global Surveillance
  - Abusive Marketing
  - Phishing and Spamming
  - Targeted Deanonymisation

Attacker Model

- Attacker knows second graph $\mathcal{S}_{AUX}$
- Probabilistic Auxiliary Information
  - Edge attribute “friendship type”
  - does not cover xor-attributes well

Seed identification

- Assumption:
  - attacker knows of k-clique in $\mathcal{S}_{SAN}$, $\mathcal{S}_{AUX}$
- Input:
  1. target graph ($\mathcal{S}_{SAN}$)
  2. k seed nodes in $\mathcal{S}_{SAN}$
  3. k node degrees
  4. $\binom{k}{2}$ common neighbour counts
- Output: matching clique in $\mathcal{S}_{SAN}$
- error factor $1 + \epsilon$ on degree and common neighbour
- return after first match improves efficiency

Propagation

- Input:
  - $G_1 = (V_1, E_1), G_2 = (V_2, E_2)$
  - partial seed $\mu_{\mathcal{S}} : V_1 \to V_2$
- Output: mapping μ
- Iteration:
  1. take arbitrary unmapped $u \in V_1$
  2. compute for each $v \in V_2$: |neigh of u mapped to neigh of v|
  3. if $> thresh$ include (u, v) in mapping
- Additional Heuristics:
  - Minimum eccentricity (of score)
  - Edge directionality (sum in- and outgoing)
  - Bias of high-degree nodes (divide score by $\sqrt{deg(v)}$)
  - Revisit nodes (more errors in empty mapping)
  - Reverse match (compute with reversed inputs)

Experiments I

- 100K-nodes graph from real social network
- random perturbation of auxiliary
- Number of seeds crucial for propagation
  - low threshold e.g. 30
  - overlap node: 25%, edge: 50% $\Rightarrow\sim 75\%$ re-identified
- Propagation robust against perturbation

Experiments II

- Twitter, Flickr, LiveJournal crawlings of 2007/'08

Network	Nodes	Edges	Av. Deg.
Twitter	224K	8.5M	37.7
Flickr	3.3M	53M	32.2

[Narayanan/Shmatikov 2009]

  - 150 nodes seed
  - name-based ground truth mapping (error $< 5\%$)
  - 30.8% re-identified, 12.1% incorrect, 57% not identified
  - 41% of incorrect are within dist 1 of true mapping
  - 55% of incorrect are within similar geo-location
  - 27% of incorrect are completely erroneous

Literatur I

- J. Kogan, *Introduction to Clustering Large and High-Dimensional Data*. Cambridge University Press, 2007.
- A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, ser. Prentice Hall Advanced Reference Series : Computer Science. Prentice Hall, March 1988.
- A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proceedings of the 30th IEEE Symposium on Security and Privacy (S&P 2009), 17-20 May, Oakland, California, USA*. IEEE Computer Society, 2009, pp. 173–187. [Online]. Available: http://www.cs.utexas.edu/~shmat/shmat_oak09.pdf