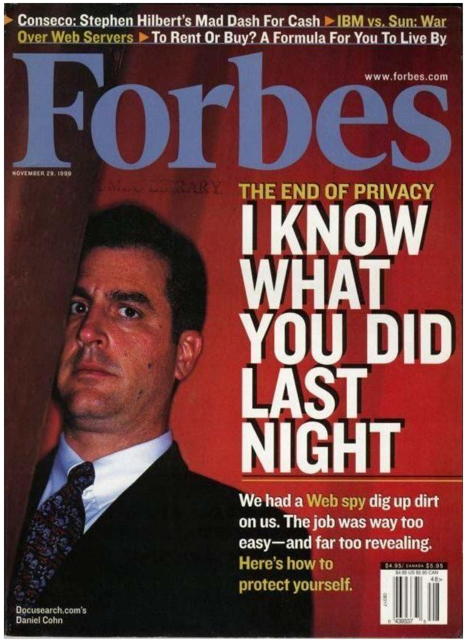
- Towards identity-anonymization on graphs

  K. Liu & E. Terzi, SIGMOD 2008

Growing Privacy Concerns

- Person specific information is being routinely collected.



"Detailed information on an individual's credit, health, and financial status, on characteristic purchasing patterns, and on other personal preferences is routinely recorded and analyzed by a variety of governmental and commercial organizations."

- M. J. Cronin, "e-Privacy?" Hoover Digest, 2000.

Proliferation of Graph Data

http://www.touchgraph.com/

Privacy breaches on graph data

- Identity disclosure
  - Identity of individuals associated with nodes is disclosed
- Link disclosure
  - Relationships between individuals are disclosed
- Content disclosure
  - Attribute data associated with a node is disclosed

Identity anonymization on graphs

- Question
  - How to share a network in a manner that permits useful analysis without disclosing the identity of the individuals involved?
- Observations
  - Simply removing the identifying information of the nodes before publishing the actual graph does not guarantee identity anonymization.

    L. Backstrom, C. Dwork, and J. Kleinberg, “Wherefore art thou R3579X?: Anonymized social netwoks, hidden patterns, and structural steganography,” In WWW 2007.

    J. Kleinberg, “Challenges in Social Network Data: Processes, Privacy and Paradoxes, ” KDD 2007 Keynote Talk.
- Can we borrow ideas from ***k***-anonymity?

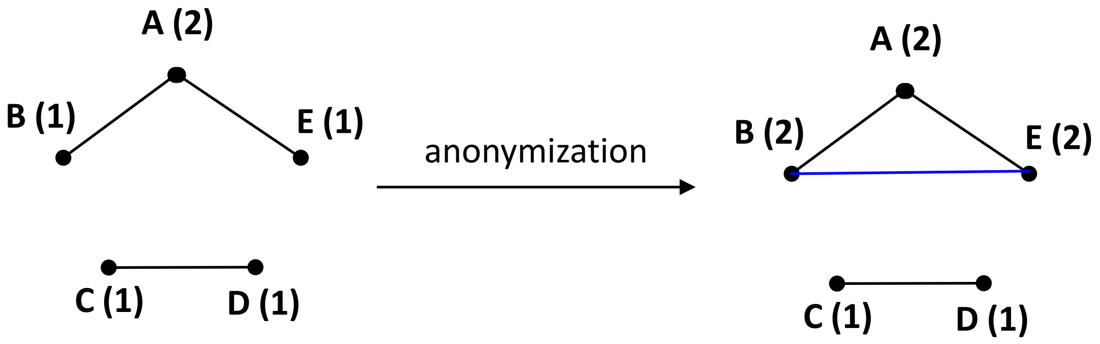
What if you want to prevent the following from happening

- Assume that adversary **A** knows that **B** has 327 connections in a social network!

- If the graph is released by removing the identity of the nodes
  - **A** can find all nodes that have degree 327
  - If there is only one node with degree 327, **A** can identify this node as being **B**.

Privacy model

[k-degree anonymity] A graph ***G(V, E)*** is ***k**-degree anonymous* if every node in ***V*** has the same degree as ***k-1*** other nodes in ***V***.



[Properties] It prevents the re-identification of individuals by adversaries with *a priori* knowledge of the degree of certain nodes.

Problem Definition

Given a graph ***G(V, E)*** and an integer ***k***, modify ***G*** via a **minimal** set of edge addition or deletion operations to construct a new graph ***G'(V', E')*** such that

1) ***G'*** is ***k***-degree anonymous;

2) ***V' = V***;

3) The ***symmetric difference*** of ***G*** and ***G'*** is as small as possible

- Symmetric difference between graphs ***G(V,E)*** and ***G'(V,E')*** :

$$\mathrm{SymDiff}(G', G) = (E' \setminus E) \cup (E \setminus E')$$

GraphAnonymization algorithm

Input: Graph ***G*** with degree sequence ***d***, integer ***k***

Output: ***k***-degree anonymous graph ***G'***

[**Degree Sequence Anonymization**]:

- Contruct an anonymized degree sequence ***d'*** from the original degree sequence ***d***

[**Graph Construction**]:

[**Construct**]: Given degree sequence ***d'***, construct a new graph $G^0(V, E^0)$ such that the degree sequence of G^0 is ***d'***

[**Transform**]: Transform $G^0(V, E^0)$ to $G'(V, E')$ so that ***SymDiff***(G',G) is minimized.

Degree-sequence anonymization

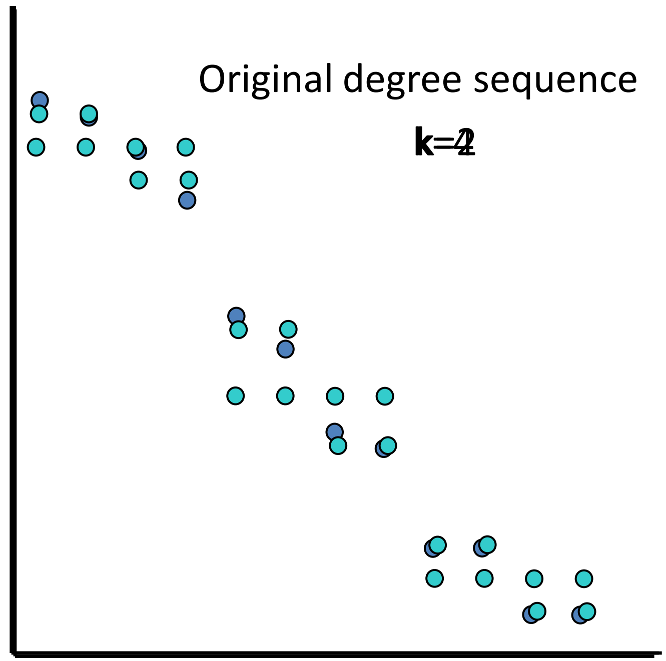
[k-anonymous sequence] A sequence of integers ***d*** is ***k**-anonymous* if every distinct element value in ***d*** appears at least ***k*** times.

[100,100, 100, 98, 98,15,15,15]

[degree-sequence anonymization] Given degree sequence ***d***, and integer ***k***, construct ***k***-anonymous sequence ***d'*** such that ***||d'-d||*** **is minimized**

Increase/decrease of degrees correspond to additions/deletions of edges

Algorithm for degree-sequence anonymization



DP for degree-sequence anonymization

- ***d*** *(1)* ≥ ***d*** *(2)* ≥… ≥ ***d*** *(i)* ≥… ≥ ***d*** *(n)* : original degree sequence.
- ***d'*** *(1)* ≥ ***d'*** *(2)* ≥…≥ ***d'*** *(i)* ≥…≥ ***d'*** *(n)* : k-anonymized degree sequence.
- ***C(i, j)*** : anonymization cost when all nodes *i, i+1, …, j* are put in the same anonymized group, i.e.,

$$C(i, j) = \sum_{\ell=i}^{j} \left(d(i) - d^*\right)$$

- ***DA(1, n)*** : the optimal degree-sequence anonymization cost
- Dynamic Programming with **O(n²)**

$$DA(1,i) = \min_{k \le t \le i-k} \{DA(1,t) + C(t+1,i)\}$$

- Dynamic Programming with **O(nk)**

$$DA(1,i) = \min_{\max\{k, i-2k+1\} \le t \le i-k} \{DA(1,t) + C(t+1,i)\}$$

- Dynamic Programming can be done in **O(n)** with some additional bookkeeping

GraphAnonymization algorithm

Input: Graph ***G*** with degree sequence ***d***, integer ***k***

Output: ***k***-degree anonymous graph ***G'***

[**Degree Sequence Anonymization**]:

- Contruct an anonymized degree sequence ***d'*** from the original degree sequence ***d***

[**Graph Construction**]:

[**Construct**]: Given degree sequence ***d'***, construct a new graph ***$G^0(V, E^0)$*** such that the degree sequence of ***G^0*** is ***d'***

[**Transform**]: Transform ***$G^0(V, E^0)$*** to ***G'(V, E')*** so that ***SymDiff****(G',G)* is minimized.

Are all degree sequences realizable?

- A degree sequence ***d*** is **realizable** if there exists a simple undirected graph with nodes having degree sequence ***d.***

- Not all vectors of integers are realizable degree sequences
  - d = {4,2,2,2,1} ?

- How can we decide?

Realizability of degree sequences

[**Erdös and Gallai**] A degree sequence ***d*** with ***d***(*1*) ≥ ***d***(*2*) ≥… ≥ ***d***(*i*) ≥… ≥ ***d***(*n*) and Σ***d****(i)* even, is realizable if and only if

$$\sum_{i=1}^{l} \boldsymbol{d}(i) \leq l(l-1) + \sum_{i=l+1}^{n} \min\{l, \boldsymbol{d}(i)\}, \text{ for every } 1 \leq l \leq n-1.$$

Input: Degree sequence ***d'***
Output: Graph ***$G^0(V, E^0)$*** with degree sequence ***d'*** or ***NO!***

→If the degree sequence ***d'*** is NOT realizable?

- Convert it into a realizable and ***k***-anonymous degree sequence

GraphAnonymization algorithm

Input: Graph ***G*** with degree sequence ***d***, integer ***k***

Output: ***k***-degree anonymous graph ***G'***

[**Degree Sequence Anonymization**]:

- Contruct an anonymized degree sequence ***d'*** from the original degree sequence ***d***

[**Graph Construction**]:

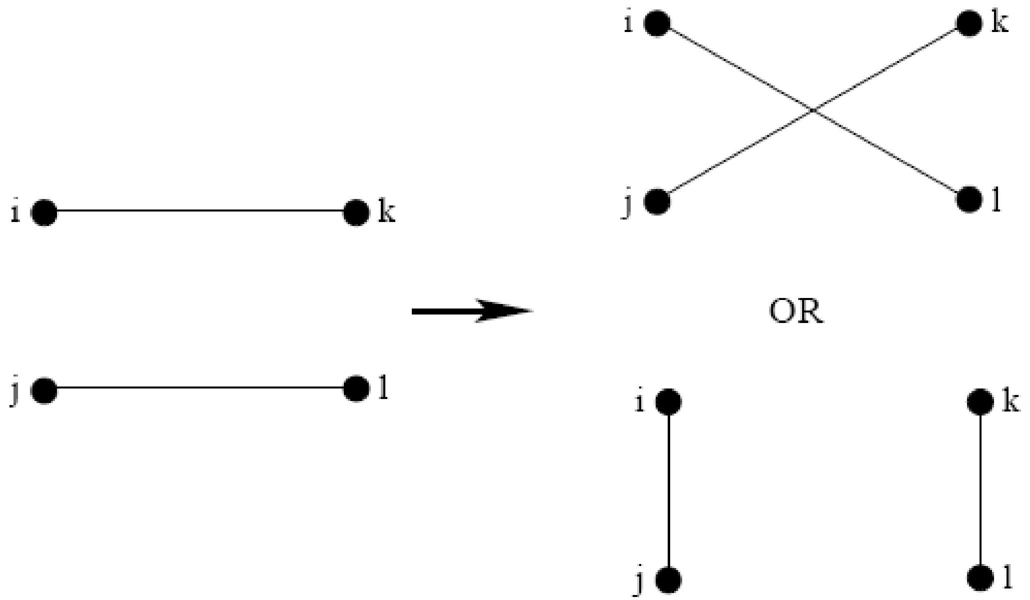
[**Construct**]: Given degree sequence ***d'***, construct a new graph $G^0(V, E^0)$ such that the degree sequence of G^0 is ***d'***

[**Transform**]: Transform $G^0(V, E^0)$ to $G'(V, E')$ so that ***SymDiff****(G',G)* is minimized.

Graph-transformation algorithm

- GreedySwap transforms $G^0 = (V, E^0)$ into $G'(V, E')$ with the same degree sequence d', and min symmetric difference $SymDiff(G',G)$.
- GreedySwap is a greedy heuristic with several iterations.
- At each step, GreedySwap swaps a pair of edges to make the graph more similar to the original graph G, while leaving the nodes' degrees intact.

Valid swappable pairs of edges



A swap is ***valid*** if the resulting graph is simple

GreedySwap algorithm

Input: A pliable graph $G^0(V, E^0)$, fixed graph $G(V,E)$

Output: Graph $G'(V, E')$ with the same degree sequence as $G^0(V,E^0)$

---

i=0

Repeat

find the valid swap in G^i that most reduces its symmetric difference with G , and form graph G^{i+1}

i++

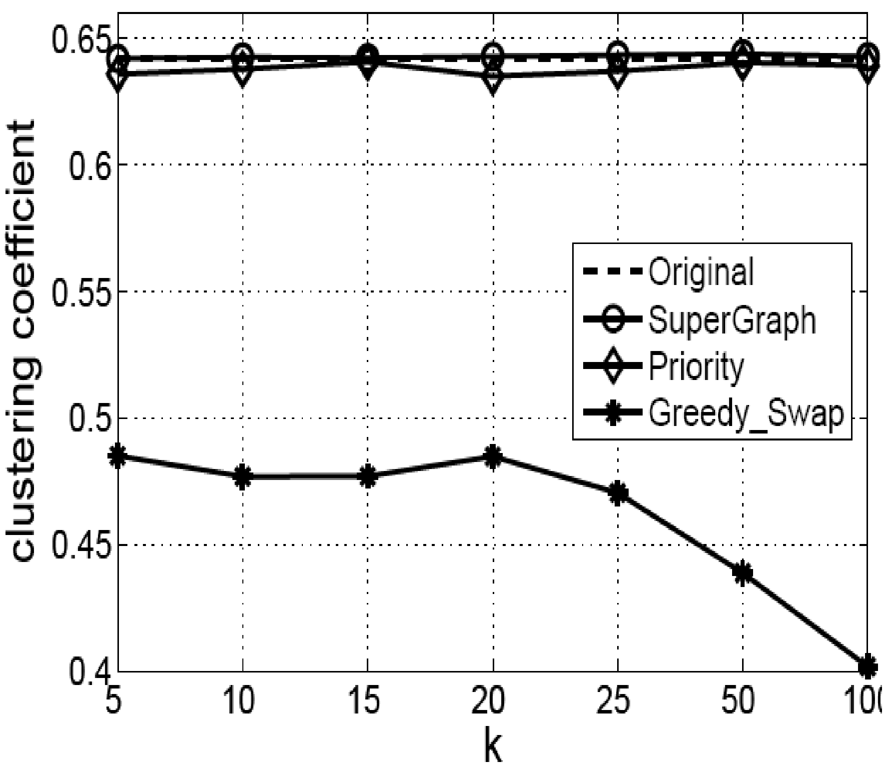
Experiments

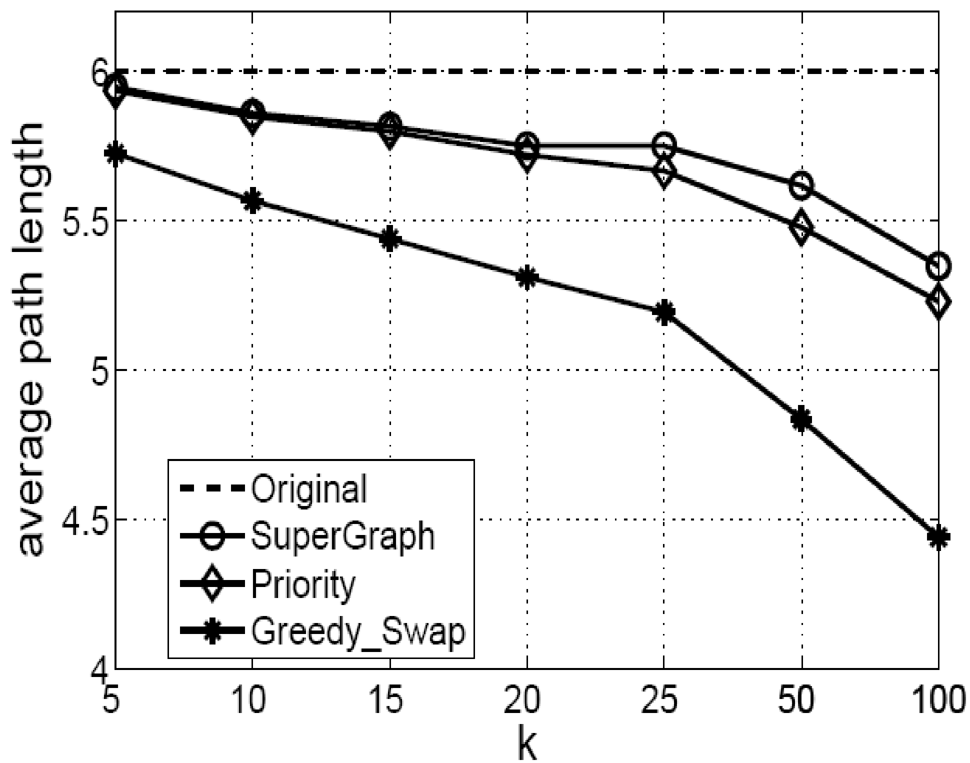
- **Datasets:** Co-authors, Enron emails, powergrid, Erdos-Renyi, small-world and power-law graphs

- **Goal:** degree-anonymization does not destroy the structure of the graph
  - Average path length
  - Clustering coefficient
  - Exponent of power-law distribution

Experiments: Clustering coefficient and Avg Path Length

- **Co-author** dataset
- APL and CC do not change dramatically even for large values of ***k***





Experiments: Edge intersections

Edge intersection achieved by the GreedySwap algorithm for different datasets.

Parenthesis value indicates the original value of edge intersection

Synthetic datasets	
Small world graphs*	0.99 (0.01)
Random graphs	0.99 (0.01)
Power law graphs**	0.93 (0.04)
Real datasets	
Enron	0.95 (0.16)
Powergrid	0.97 (0.01)
Co-authors	0.91(0.01)

(*) L. Barabasi and R. Albert: Emergence of scaling in random networks. *Science 1999.*

(**) Watts, D. J. Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology* 1999

Privacy in transaction data

Voter Registration List

Name	DOB	Sex	Zipcode
Andre	1/21/76	Male	53715
Beth	1/10/81	Female	55410
Carol	10/1/44	Female	90210
Dan	2/21/84	Male	02174
Ellen	4/19/72	Female	02237

Patient Records

ID	DOB	Sex	Zipcode	Disease
1	1/21/76	Male	53715	Flu
2	1/21/76	Male	53703	Broken Arm
3	9/1/86	Male	53715	Bronchitis
4	4/13/86	Female	53715	Hepatitis
5	2/28/86	Female	53708	Flu
6	2/28/86	Female	53708	HIV

Data *De-Identification*

- **Identifiers** typically removed
  - e.g., Name and Social Security #
- Threat of re-identification by linking public data sets using other attributes
  - e.g., DOB, Sex, and Zipcode
- Refer to the set of attributes available externally as the **quasi-identifier**
  - Assume known based on the domain

k-Anonymity

- Intuitive means of protecting ***identity***
- Single published table ***T***
- Generalize / suppress quasi-identifier values so no individual uniquely identified from a group smaller than ***k***
  - Each group of records with identical quasi-identifier values is a **QI-group**
  - Table ***T*** is ***k*-anonymous** if the size of each QI-group is at least ***k***.

Example

Voter Registration List

Name	DOB	Sex	Zipcode
Andre	1/21/76	Male	53715
Beth	1/10/81	Female	55410
Carol	10/1/44	Female	90210
Dan	2/21/84	Male	02174
Ellen	4/19/72	Female	02237

Patient Records

ID	DOB	Sex	Zipcode	Disease
1	1/21/76	Male	537**	Flu
2	1/21/76	Male	537**	Broken Arm
3	1986	*	53715	Bronchitis
4	1986	*	53715	Hepatitis
5	2/28/86	Female	53708	Flu
6	2/28/86	Female	53708	HIV

Example

Voter Registration List

Name	DOB	Sex	Zipcode
Andre	1/21/76	Male	53715
Beth	1/10/81	Female	55410
Carol	10/1/44	Female	90210
Dan	2/21/84	Male	02174
Ellen	4/19/72	Female	02237

Patient Records

ID	DOB	Sex	Zipcode	Disease
1	1/21/76	Male	537**	**Hepatitis**
2	1/21/76	Male	537**	**Hepatitis**
3	1986	*	53715	Bronchitis
4	1986	*	53715	Hepatitis
5	2/28/86	Female	53708	Flu
6	2/28/86	Female	53708	HIV

Competing Goals

- ***Privacy*** vs. ***Utility***
- Released data should be as useful as possible, while respecting privacy constraints.

Key Questions

How should we manipulate published data to satisfy *k*-anonymity? Preserve utility?

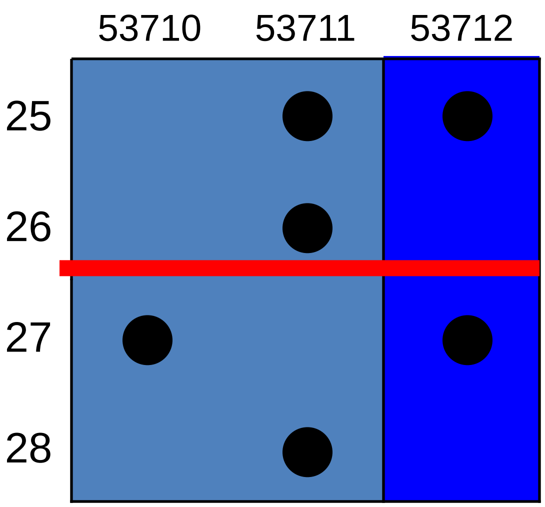
Single-Dimensional Global Recoding

- Each quasi-identifier attribute $\boldsymbol{X_i}$ has some domain of unique values ($\boldsymbol{D_{Xi}}$)
- Map each $\boldsymbol{D_{Xi}}$ to "generalized" set of values

Single-Dimensional Global Recoding

- Divide each quasi-identifier domain (individually) into ***ranges***

k=2



Age	Zipcode
[25-28]	[53710-53711]
[25-28]	[53710-53711]
[25-28]	[53710-53711]
[25-28]	[53710-53711]
[25-28]	53712
[25-28]	53712

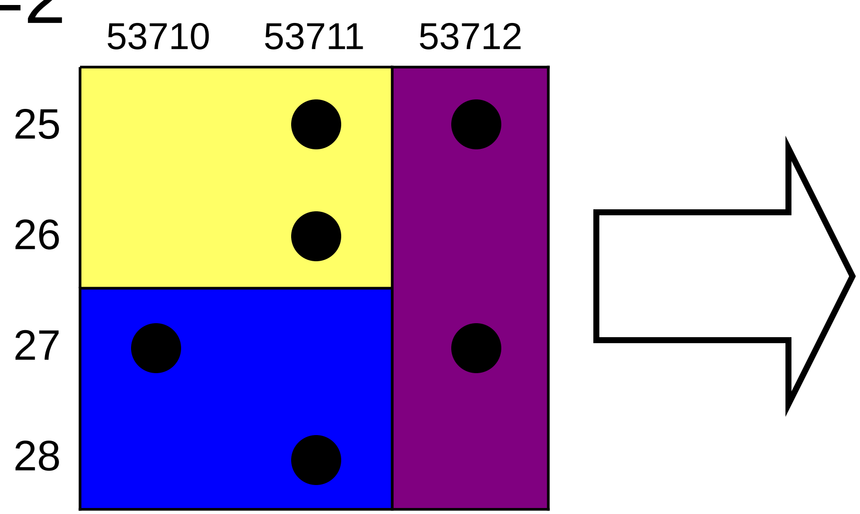
Multidimensional Global Recoding

- ***Flexible Alternative…***
  - Map $\boldsymbol{D_{X1} \times \ldots \times D_{Xn}}$ to “generalized” set of vector values
  - Every single-dimensional recoding can be expressed as a multidimensional recoding

Multidimensional Global Recoding

- Set of non-overlapping hyper-rectangular ***regions*** covering domain space

k=2



Age	Zipcode
[25-26]	[53710-53711]
[25-26]	[53710-53711]
[27-28]	[53710-53711]
[27-28]	[53710-53711]
[25-28]	53712
[25-28]	53712