A SURVEY ON PRIVACY PRESERVATION TECHNIQUES FOR DATA CLUSTERING K-MEANS OVER LARGE-SCALE DATASET

C S shriitha
Mtech, dept. of computer science and engineering
Gokaraju Rangaraju Engineering College
Hyderabad, India

Dr A Sai Hanuman
Professor of computer science and engineering
Gokaraju Rangaraju Engineering College
Hyderabad, India

Abstract: Cloud computing supports different handling of Big-Data applications in such divisions like human services and Sports and so on. Data sets like electronic wellbeing records is regularly contain protection touchy data, which achieves security concerns possibly if the data is discharged/shared to outsiders in cloud. A functional and broadly received procedure for protection safeguarding is to anonymize information by means of speculation to fulfill a given security demonstrates. In this paper, we propose a viable security safeguarding K-means grouping plan that can be effectively outsourced to cloud servers. The present work permits cloud servers to perform bunching specifically finished encoded datasets, while achieving comparable computational complexity and accuracy compared with clustering’s over unencrypted ones. In addition to existing techniques, MapReduce approach also combined in this paper, which makes this work greatly appropriate for MapReduce condition. Differentially security approach ensures the results of questions to a database, which will expand the versatility and time proficiency over existing methodologies.

Keywords: Cloud Computing, Big data, MapReduce, Data Anonymization, K-means algorithm

I. INTRODUCTION

Big Data and Cloud Computing, a critical effect on IT industry and research groups where expansive measure of information can store and recover [11, 12]. Distributed computing is an imaginative administration mode. It empowers clients to get practically boundless processing power and copious an assortment of data administrations from web. They are disseminated processing, parallel registering and lattice computational advancement. This sort of new example eludes the incorporation and extension to the IT foundation, through the system to the required assets (equipment, stage, and programming), virtual mix into a dependable and superior processing stage. In distributed computing, all clients’ information are put away in the cloud resources Nodes [2, 13, 1, 7, 14]. The outcomes disseminate to the client through the system when the client required.

The majority of the mechanical information put away in cloud computing, however can't anticipate all put away information more likely than not secured, thus a large proportion of cloud information are encoded. Significantly more encryption calculation imagined, touchy data can spill if that one key is released in this way, less secure. The vast majority of the encryption key is overseen by cloud suppliers, so suppliers may break all data. "Two level Encryption" for that we client direct congruential generator and DES algorithm, all put away data have two classification, one for look record another protection table. Hunt list contain just accessible catchphrases.

Security table are kept up by organize administrator that contain one of a kind encryption keys for all patient. These key just give approved demand that implies patient can set direction for get to our key. For the most part for the security protection the information anonymization method has been utilized. Information anonymization is to shroud the touchy data so the security for an individual is exceedingly safeguarded. Adaptability and productivity challenges are by the 3V's they are Volume, Velocity and Variety. For the cell age the nearby recoding ideas are utilized, where the information is gathered as set of cells and anonymize each record independently.

II. LITERATURE SURVEY

As of late clustering techniques has been enhanced or upgraded to accomplish a protection safeguarding in neighborhood recoding anonymization.[1] From the utility security conservation viewpoint the nearby recoding anonymization has been examined. It likewise utilizes the best down partner and a base up avaricious approach are as one pit-forward in view of the bunch measure, the agglomerative grouping procedure and disruptive bunching systems get enhanced.[2] Data security safeguarding has been examined widely, existing methodologies for neighborhood recoding anonymization and models for protection are evaluated quickly. Likewise, the exploration for adaptability issues in existing anonymization approaches are reviewed in the blink of an eye.

To address the neighborhood recoding anonymization as the k-means bunching issue where the group size ought not be not as much as k to accomplish k-obscurity. For that the straightforward ravenious calculation are used.[3] For the various leveled properties, KACA(k-Anonymization by Clustering in Attribute Hierarchies) algorithms are proposed for irregularity issue of nearby recoding anonymization in data.[4] Existing grouping approach for neighborhood recoding anonymization fundamentally focus on record
linkage assaults principally under the k-namelessness security demonstrate, with no significance to protection ruptures acquired by delicate quality linkage. Moderately propose a consistent factor guess calculation for two grouping based anonymization issue, that is r-GATHER and r-CELLULAR CLUSTERING, here the habitats for bunches are distributed without speculation or suppression.[5] Based on Mondrian calculation a best down dividing approach get proposed [6] to safeguard certain trait linkage assaults by practice informational collections to achieve (α,k) obscurity. By part traits and qualities, the information utility of the resultant mysterious information is vigorously impacted, while nearby recoding does not include such factors. This approach enhances bunching to finish nearby recoding since, it is a characteristic and powerful approach to anonymize informational collections at a cell level. In some protection models, the certainty of partner a semi identifier to a touchy incentive to be not as much as a client indicated edge by (α,k)anonymity [5]. Difference control [7] tended to the versatility issue of multi-dimensional anonymization plot [6] by means of presenting adaptable choice trees and examining systems. For accomplishing high productivity, a R-tree list based methodologies get proposed by building an uncommon file over informational collections. Through utilizing MapReduce worldview the issue for the sub-tree plot in huge information situation is tended to in light of our past work [8, 9]. Fundamentally these methodologies are utilized to keep the checks from spilling excessively data. In [16] a wavelet change is connected to the information and the clamor is included the recurrence space. In [15, 17] the histogram canisters are changed in accordance with the genuine information. In [18] the differential security of traits whose space is requested and has direct to extensive cardinality like numerical characteristics, the properties are spoken to as tree, to expand the exactness of answers to check questions they get disintegrated. Differential security is like [16] where the first information will change, yet our proposed framework will manage both requested and not requested qualities. A down to earth strategy [19] demonstrated that a fulfillment of a casual type of differential security by a compelled k-anonymization continued by irregular inspecting.[20] On the contrary, to lessen the data misfortune caused by standard differential protection k-anonymization is utilized.

III. PRELIMINARIES AND PROBLEM ANALYSIS

Anonymization Scheme for Local-Recoding

Local-recoding is the type of cell-speculation. It is the one of the plan for separate the information as cell. Different plans are full-space, sub-tree and multi-dimensional anonymization. Nearby recoding sums up the informational index as cell level, where the worldwide recoding sums up the informational collection as the area level. Regularly, neighborhood recoding limits the information contortion by data cleansing and along these lines create preferable information utility over worldwide recoding. As a rule, anonymization is for protection conservation.

Basics of MapReduce

One of the huge scale information handling ideal models is MapReduce where, it has been widely looked into and successively received for huge information applications as of late [21]. MapReduce is more adaptable and practical because of notable highlights and attributes of distributed computing, A correct case for MapReduce is Amazon Elastic MapReduce benefit. Essentially, MapReduce work comprise of two segments Map and Reduce where, it is named as key esteem combine (key, esteem). Formally, Map work is named as Map: (k1,v1)→(k2,v2), i.e., the guide work take (k1,v1) as an information and deliver another key-esteem combine (k2,v2). Thus, Reduce capacity can be indicated as Reduce (k2,list(v2))→(k3,v3), that is Reduce work takes contribution as (k2,list(v2)) and create yield as (k3,v3). At long last, yield for MapReduce work is consider as (k3, v3). Both Map and Reduce work are determined by the client as per their particular applications.

Motivation and Problem Analysis

In this area, the issues distinguished from the current methodologies are dissected for protection saving and adaptability. Some significance will provided for the neighborhood recoding procedure for the record linkage assaults over the informational collections. For the adaptability reason the t-grouping approach have been utilized to parcel the informational collections into guide task toward shape a bunch. By framing a group, from that whatever remains of the records will relegate into this bunches. Also, ε-differential security strategy is utilized to ensure the results of inquiries to a database. Consecutively, plan a legitimate map reduce occupations for complex applications for the most part for the parallelized issue and for arrange traffics among information nodes.

IV. TWO-PHASE PRIVATE CLUSTERING USING MAPREDUCE

Design of Two-phase clustering

For the portrayal of group task the t-progenitor strategy utilized. In t-progenitor calculation each straight out semi identifier is the least basic predecessor of the first incentive in the group. In progenitor record middle of the first esteem will be the numerical semi identifier. In grouping issue for anonymization, t-predecessors bunching is great. Through the separation estimation, the separation between information records and precursors will compute. For versatility viewpoint, point-task strategies are perfect for neighborhood recoding anonymization in MapReduce. Point bunches are utilized to pick an arrangement of information records to shape a group, from that whatever is left of the records will dole out into these bunches. Point task will rehash until the point that the condition fulfilled. Be that as it may, for the extensive arrangement of information records under perceptions, the size will be 1/k of a unique informational collection.

One of the issues here is, while point task process, the measure of the group will wild. At the point when the informational collection has high skewness, the group can surpass the upper bound 2k-1 or be not as much as k. In the main stage, point task grouping strategy used to parcel the first informational collection into t-bunches. A bunch created in the principal stage is named α-group in the second
stage, ε-differential security to ensure the results of inquiries to a database.

V. SURVEY APPROCHES

Algorithm1: Design of Two-Phase Clustering
Input: Data set B, anonymity parameter k
Output: Anonymous data set B*
1. Run the t-ancestor clustering algorithm on B, get a set of α-clusters: $C^α=\{C_1^α,\ldots,C_n^α\}$.
2. For each α-cluster $C_i^α \in C^α$, $1 \leq i \leq t$; run ε-differentialprivacy algorithm
   Let $S(t)$ be an ε-differentially private sanitizer
   $∀ \ y \in \text{data set TA(Y)} \text{for } R=1 \text{ to } n \ \text{do} \ y \leftarrow Sτ(Qr(\bar{y}))$
   End for
3. Return $\text{V}_ε$
4. For each cluster $C_i \in C$ , where $C=U_{i=1}^t C_i$ .
generate $C_i^{*}$ by replacing each attribute value with a generalone.
5. Generate $B^{*} = U_{i=1}^t C_i^{*}$, where $m_\text{t} = \sum_{i=1}^{t} m_1$

Technique for producing the differentially private informational index X. Let X is an informational index with m numerical traits. The initial step to run the t-ancestor clustering algorithm on the informational collection B. The area of X contains all the conceivable esteem that bode well for the space isn't characterized by the genuine records in X however by the arrangement of qualities that bode well for each trait and by the connection between characteristics.

Partitioned Data by t-ancestor clustering
At first the t-ancestors, in light of the point assignments the t-records are chosen as seeds. Such determination utilizing the point task will make the t-records impacts the nature of bunching to certain expands. By picking the records for far from each other will make the arrangement of seeds great. Utilizing map-Reduce work, the seeds get chose as seed determination which yields an arrangement of seeds: $S_1=\{R_1, \ldots, R_t\}$. The guide and lessen elements of seed determination which yields an arrangement of seeds:

```
Input: B, parameter t, thresholds τ,φ
Output: α-clusters $C^α=\{C_1^α,\ldots,C_n^α\}$
1: Run job Seedselection; get initial seeds $S_1^α$; $i\leftarrow 0$;
2: Run job Ancestorupdate; get ancestors $S_{(i+1)}^α$; $i\leftarrow i+1$;
While $d(S_{(i)}^α, S_{(i+1)}^α)\geq τ$ and $i\leq φ$, repeat step 2;
3: Return α-clusters with ancestors $S_{(i)}^α$.
```

In each round of while-loop in algorithm3, to full-fill the expectation an maximization steps, a MapReduce jobnamed as Ancestor Update is designed. Initially, the map function is responsible for point assignment in the expectation (E) step, while the reduce function is responsible for computation of ancestors (M) step. Map and reduce functions are described in algorithm4. Two subroutines Average() and Ancestor() are utilized to calculate the medians of numerical attributes and ancestors of categorical attributes,进而Reduce function. One Reduce function can process more than one α-clusters in sequence if t is large enough. Reduce function will scalably with setting t.

Algorithm2: Seed Selection Map and Reduce
Input: Data record R, ReB
Output: A set of Seeds $S_i=\{R_1, \ldots, R_t\}$
Map: Generate a Random value
Random, where $0 \leq \text{rand} \leq 1$; if $\text{Random} \leq N/|B|$, emit (1,R)
Reduce: 1. Select a random record R from list(R), $S_1\leftarrow R$;
2. While $|S_i|<t$;
Find R ∈ list(R) that maximizes $\min_R cS_i d(R,R_1)$;
$S_i\leftarrow R$
3. Emit (null,$S_i$)

The t-ancestor algorithm takes after an emphasis refinement system for each datum records. Principally two stages are taken after for each cycle in particular expectation (E) and Maximization (M) [10]. In desire (E) and information records are doled out to their closest progenitor and constitute a α-bunch. In Maximization (M) step, the computation is performed to each record in bunch for the predecessor of a α-group. In E-step, the new arrangement of predecessors has been utilized as a part of next round. It is normal that the cycle joins, ie, after a limited number of records, the assignments have never again changes. The separation estimations are utilized as a part of t-grouping, 1) the distinction of progenitors between two nonstop adjusts of cycle land at predefined limit.

First stopping criteria is determined by $d(S_{(i)}^α, S_{(i+1)}^α)<τ$, where $τ$ is a predefined threshold. Let $\text{V}_1$ denote the maximum number of iteration rounds is predefined ifany one of the above criteria gets satisfied means, the t-ancestorclustering algorithm get stops.

Algorithm3: t-Ancestor Clustering Approach
Input: Data set B, parameter t, thresholds τ,φ
Output: - α-clusters $C^α=\{C_1^α,\ldots,C_n^α\}$
1: Run job Seedselection; get initial seeds $S_1^α$; $i\leftarrow 0$;
2: Run job Ancestorupdate; get ancestors $S_{(i+1)}^α$; $i\leftarrow i+1$;
While $d(S_{(i)}^α, S_{(i+1)}^α)\geq τ$ and $i\leq φ$, repeat step 2;
3: Return α-clusters with ancestors $S_{(i)}^α$.

Algorithm4: Ancestor Update Map
Input: Data Record R, ReB; Seeds of round $I,S_{(i)}^α=\{R_1,\ldots, R_t\}$
Output:Seeds of round $I$, $S_{(i+1)}^α=\{R_{(i+1)}^1,\ldots, R_{(i+1)}^t\}$
Map: 1. dmin$\leftarrow +\infty$;
2. for j: 1 to t
If$d(R,R_j)<d_{\text{min}}$, then $d_{\text{min}}\leftarrow d(R,R_j)$ and $j\leftarrow j$;
3. Emit $(j_{\text{min}}, R)$
Reduce: 1. for I: 1 to n$^o$
If attr$^{v_i}$ is numerical, then $V_i\leftarrow$ Average (list(R),I);
Else $V_i\leftarrow$ Ancestor (list(R);I);
2. Emit $(j,R_j^{(i+1)}=\langle V_1,\ldots, V_n^{(i)}\rangle)$.
Differential Privacy Data Sets Through K-anonymization

For numerical attributes, the generation of the ε-differential privacy data set \(Y_\varepsilon\) as described in previous methods.

Give Y a chance to be a dataset with n numerical traits: \(a_1\) … \(a_n\). The first step is to develop \(Y_\varepsilon\) is to produce k-mysterious informational index \(\gamma\) by means of a t-progenitor bunching calculation. We create q-by-questioning Y with \(\text{Ir}(Y)\), for \(r = 1 \text{ to } n\). On the off chance that the reactions to the questions \(\text{Ir}(r)\) fulfill ε-differential protection, at that point each inquiry alludes to various record. \(Y_\varepsilon\) likewise fulfill ε-differential security. By giving a differentially private reaction to the questions for all qualities in each record, the differentially private informational collection \(Y_\varepsilon\) is created.

By developing k-known informational index \(\gamma\), the terms are assembled in the k-people. Presently, the defensively of the questions \(\text{Ir}(\gamma)\) used to develop \(Y_\varepsilon\) mirrors the impact that changing a solitary record in Y has on the gatherings of k-records in \(\gamma\). Each record in \(\gamma\) relies upon k records in Y is prompts the diminished affectability of the arrangement of inquiries \(\text{Ir}(\gamma)\) is littler than the affectability of the arrangement of questions \(\text{Ir}(Y)\), for \(r = 1 \text{ to } n\). We seen that the defensively of every individual question \(\text{Ir}(\gamma)\) is upper limited by \(\Delta \text{Ir}(Y)/k\). Having \(n/k\) diverse inquiries \(\gamma\) the affectability of \(\text{Ir}(\gamma)\) for \(r \in \gamma\) is upper limited by \(n/k \times \Delta \text{Ir}(Y)/k\). By changing the group estimate k, the estimations of \(n/k \times \Delta \text{Ir}(Y)/k\) will be littler than \(\Delta \text{Ir}(Y)\). Expanding the group size will lessens the commitment of each record to the bunch centroid and it decreases the quantity of created bunches. By utilizing the t-grouping calculation for point task, the adaptability issue will get lessen.

Algorithm 5: Generation of a ε-Differential Privacy Data Set \(Y_\varepsilon\)and Y via t-ancestor clustering

Let Y be an original data set with n records
Let TA be an t-ancestor clustering algorithm with minimal cluster size k
Let \(S\alpha()\) be an ε-differentially private sanitizer
Let \(\text{Ir}()\) be the query for attributes of the r-th record
\[ \gamma \leftarrow \text{Partitioned data set } TA(Y) \]
for \(R = 1 \text{ to } n \) do
\[ \gamma \varepsilon \leftarrow S\alpha(\text{Ir}(\gamma)) \]
insert \(\gamma \varepsilon\) into \(Y_\varepsilon\)
End for
Return \(Y_\varepsilon\)

Give Y a chance to be an informational collection with n clear cut traits \(A_1\) … \(A_m\). The difficulties respect the meaning of \(\text{Dom}(Y)\). The universe of each clear cut characterizations can characterize by augmentation, posting every single conceivable esteem. This level rundown from characteristics can characterize by augmentation, posting meaning of \(\text{Dom}(Y)\). The universe of each clear cut \(A_i\) can be characterized as the requested mix of estimations of each \(\text{Dom}(A_i)\) as displayed in their scientific categorizations \(\tau(A_1)\) … \(\tau(A_m)\). A semantic separation \(\delta\) evaluates the measure of semantic contrast saw between two terms. We can characterize the separation \(d\) for the scientific categorization as, \(d: \text{Dom}(Y) \times \text{Dom}(Y)\).

\[ M(\text{Dom}(A_i), a_j) = \sum_{\text{aji} \in \text{Dom}(A_i)} \delta(a_i, a_j) \]

Here, \(\delta(\ldots, \cdot)\) is the distance between values. For each \(A_i\), one boundary \(a_{i,}^1\) of \(\text{Dom}(A_i)\) can be defined as the most marginal value of \(\text{Dom}(A_i)\), \(a_{i,}^1 = \text{argmax}_{a_{i,} \in \text{Dom}(A_i)} \text{max} \delta(a_i, a_{i,})\) other boundary \(a_{i,}^1\) can be defined as the most distancevalue from \(a_{i,}^1\) in \(\text{Dom}(A_i)\). \(a_i = \text{argmin}_{a_{i,} \in \text{Dom}(A_i)} \max \delta(a_i, a_{i,})\)

By applying the above expression, the set of attributes \(A_1\), … \(A_m\) in Y the reference point needed to define a total order according to the semantic distance can be reconstructed. Finally for a sample of \(\tau(Z(A_i))\) of a nominal attribute \(A_i\) in a certain cluster, the marginality based centroid for that cluster is defined as,

\[ \text{Centroid } (Z(A_i)) = \text{argmax}_{Z(A_i)} \min_{m_i} (Z(A_i), a_i) \]

Where \(\tau(Z(A_i))\) is the minimum taxonomy extracted from \(\text{Dom}(A_i)\) that includes all values in \(Z(A_i)\). To fulfill differential privacy to categorical attributes, the centroid computations should evaluated another one is to achieve intensity.

VI. CONCLUSION AND FUTURE WORK

In this paper, the t-clustering problem in k-anonymization has been examined in all points of view for productivity and versatility. The proposed k-means privacy approach mainly deals with a shape a bunch likewise to ensure the results of inquiries to a database. By the commitment of over two strategies for the future upgrade we intend to incorporate an arranging calculation to enhance the adaptability and security to the informational collections.

VII. REFERENCES

[1] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, “Utility based anonymization using local recoding,” in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2006, pp. 785–790.

[2] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-preserving data publishing: A survey of recent developments,” ACM Comput. Survey, vol. 42, no. 4, pp. 1–53, 2010.

[3] J. Li, R. C.-W. Wong, A. W.-C. Fu, and J. Pei, “Anonymization by local recoding in data with attribute hierarchical taxonomies,” IEEE Trans. Knowl. Data Eng., vol. 20, no. 9, pp. 1181–1194, Sep. 2008.

[4] G. Aggarwal, R. Panigrahy, T. Feder, D. Thomas, K. Kenthapadi, S. Khuller, and A. Zhu, “Achieving anonymity via clustering,” ACM Trans. Algorithms, vol. 6, no. 3, 2010.

[5] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, “(a,k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing,” in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2006, pp. 754–759.
[6] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional k-anonymity,” in Proc. 22nd Int. Conf. Data Eng., 2006, p. 25.

[7] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Workload-aware anonymization techniques for largescaledatasets,” ACM Trans. Database Syst., vol. 33, no.3, pp. 1–47, 2008.

[8] X. Zhang, L. T. Yang, C. Liu, and J. Chen, “A scalable two-phase top-down specialization approach for dataanonymization using Map reduce on cloud,” IEEE Trans. Parallel Distribute. Syst., vol. 25, no. 2, pp. 363–373, Feb.2014.

[9] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, and J. Chen, “A hybrid approach for scalable sub-treeanonymization over big data using map reduce on cloud,” J. Compute. Syst. Sci., vol. 80, no. 5, pp. 1008–1020, 2014.

[10] S. Lloyd, “Least squares quantization in PCM,” IEEE Trans. Info. Theory, vol. IT-28, no. 2, pp. 129–137, Mar.1982.

[11] S. Chaudhuri, “What next?: A half-dozen datamanagement research goals for big data and the cloud,” in Proc. 31st Symp. PrinciplesDatabase Syst., 2012, pp. 1–4.

[12] L. Wang, J. Zhan, W. Shi, and Y. Liang, “In cloud, canscientific communities benefit from the economies of scale?” IEEE Trans. Parallel Distribute. Syst., vol. 23, no.2, pp. 296–303, Feb. 2012.

[13] L. Sweeney, “K-anonymity: A model for protectingprivacy,” Int. J. Uncertainty Fuzziness, vol. 10, no. 5, pp.557–570, 2002.

[14] T. Wang, S. Meng, B. Bamba, L. Liu, and C. Pu, “A general proximity privacy principle,” in Proc. IEEE 25th Int. Conf. Data Eng., 2009, pp. 1279–1282.

[15] K.-H. Lee, Y.-J. Lee, H. Choi, Y. D. Chung, and B. Moon, “Parallel data processing with map reduce: A survey,” ACM SIGMOD Record, vol. 40, no. 4, pp. 11–20, 2012.

[16] Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G., Differential Privacy via Wavelet Transforms. In: IEEE International Conference on Data Engineering (ICDE 2012), pp. 32-43 (2012)

[17] Xiao, X., Wang, G., Gehrke, J.: Differential Privacy via Wavelet Transforms. IEEE Trans. on Knowl. and Data Eng. 23(8), pp. 1200-1214 (2010)

[18] Li, N., Yang, W., Qardaji, W.: Differentially private grids for geospatial data. In: IEEE International Conference on Data Engineering (ICDE 2013), pp. 757-768 (2013)

[19] Cormode, G., Procopiuc, C. M., Shen, E., Srivastava, D., Yu, T.: Differentially private spatial decompositions. In: IEEE International Conference on Data Engineering (ICDE 2012), pp. 20-31 (2012)

[20] Li, N., Qardaji, V., Su, D.: On sampling, anonymization, and differential privacy: Or, k-anonymization meets differential privacy. In: 7th ACM Symposium on Information, Computer and Communications Security (ASIACCS’ 2012), pages 32-33 (2012)

[21] Domingo-Ferrer, J., Sanchez, D., Rufian-Torrell, G.: Anonymization of nominal data based on semanticmarginality. Inf. Sci. 242, 35-48 (2013)