Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Sensitive Micro Data Disclosures Based on Tuple Grouping Methods

Vasudha T*
Research Scholar
Department of CSE
MIC College Of Technology
JNTUK, Andhra Pradesh.

JanakiRamaiah B*
Assoc. Professor
Department of CSE
MIC College Of Technology
JNTUK, Andhra Pradesh

Abstract: ***Data anonymization is one key aspect of Micro data disclosures as they enable policy-makers to analyze the decision outcomes of issues influencing the business there by influencing the future course of actions. Privacy is a key issue here because inappropriate disclosure of certain data assets will harm the prospects. Prior approaches of data anonymization such as generalization and bucketization (driven by k-anonymity, l-diversity) have been designed for privacy preserving micro data publishing which have several limitations like Generalization's inability to handle high dimensional data and Bucketization failure to maintain clear separation between quasi-identifying attributes and sensitive attributes prompted the development of a novel technique called Slicing, which partitions the data both horizontally and vertically. Although Slicing achieves better data utility and anonymity compared to prior techniques, its sensitive attribute disclosures are based on random grouping, which is not very effective as randomly generating the associations between column values of a bucket significantly lowers data utility. Therefore, we propose to replace random grouping with more effective tuple grouping algorithms such as Tuple Space Search algorithm based on hashing techniques. The computed and obtained sliced data from high dimensional sensitive attributes based on the proposed technique offers significant performance rise. A feasible practical implementation on dynamic data validates our claim.***

Keywords: ***Data anonymization, Data publishing, Data security, Privacy preservation.***

I. INTRODUCTION

Data mining that is sometimes also known as Knowledge Discovery Data (KDD) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is extracting the meaningful information from the large data sets such as data warehouse; Micro data contains records each of which contains information about an individual entity. Many micro data anonymization techniques have been proposed and the most popular ones are generalization with k-anonymity and bucketization with *l*diversity.For privacy in Microdatapublishing a novel technique called slicing is used that partitions the data both horizontally and vertically.

Slicing preserves better data utility thangeneralization and can be used for membershipdisclosure protection. It can handle high dimensionaldata.A bettersystem is required that can with stand high dimensionaldata handling and sensitive attributedisclosure failures. These quasi–identifiers are set of attributes are those that in combination can be linked with the external information to reidentify. There are three categories of attributes in microdata. In the case of both anonymization techniques, first identifiers are removed from the data and then partitions the tuple's into buckets.

In generalization, transforms the quasi-identifying values in each bucket into less specific and semantically constant so that tuple's in the same bucket cannot be distinguished by their QI values. One separates the SA values from the QI values by randomly permuting the SA values in the bucket in bucketization. The anonymized data consist of a set of buckets with permuted sensitive attribute values. Existing works mainly considers datasets with a single sensitive attribute while patient data consists multiple sensitive attributes such as diagnosis and treatment.

Data slicing can also be used to prevent membership disclosure and is efficient for high dimensional data and preserves better data utility. We introduce a noveldata anonymization technique called slicing toimprove the current state of the art. Data has beenpartitioned horizontally and vertically by the slicing. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes.Horizontal partitioning is done by grouping tuple's into buckets.

Slicing preservesutility because it groups highly correlated attributestogether and preserves the correlations between suchattributes.When the data setcontains QIs and one SA, bucketization has to breaktheir correlation. Slicing

can groupsome QI attributes with the SA for preserving attributecorrelations with the sensitive attribute.We present a novel technique called slicingfor privacy-preserving data publishing.

II. BACKGROUNDWORK

1. Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jian Zhang, Member, IEEE, and Ian Molloy[1].
   Several anonymization techniques, such as generalization and bucketization, slicing have been designed for privacy preserving microdata publishing.
2. Amar Paul Singh, Ms. DhanshriParihar[2].
   Most enterprises collecting data from large databases for making business decisions. This paper deals with how to provide anonymization techniques to the data.
3. M.Alphonsa, V.Anandam, D.Baswaraj[3].
   For privacy preserving and micro data publishing we proposed different methodologies.
4. Y. He and J. Naughton[4].
   The data can be represented in set valued basis in database.

III. EXISTINGSYSTEM

3.1 Generalization: Generalization is one of the commonly anonymized approaches that replace quasi-identifier values with values that are less specific but semantically consistent. All quasi-identifier values in a group would be generalized to the entire group extent in the QID space. If at least two transactions in a group have distinct values in a certain column then all information about the item in current group is lost. QID used in this process includes all possible items in the log. In order for generalization to be effective, records in the same bucket must be close to each other so that generalizing the records would not lose too much information. The data analyst has to make the uniform distribution assumption that every value in a generalized interval/set is equally possible to perform data analysis or data mining tasks on the generalized table. This significantly reduces the data utility of the generalized data.

3.2 Bucketization:

Bucketization is to partition the tuple's in T into buckets and then to separate the sensitive attribute from the non-sensitive ones by randomly permuting the sensitive attribute values within each bucket.We use bucketization as the method of constructing the published data from the original table T. We apply an independent random permutation to the column containing S-values within each bucket. The resulting set of buckets is then published. While bucketization has better data utility than generalization it has several limitations. Bucketization does not prevent membership disclosure because bucketization publishes the QI values in their original forms. Bucketization requires a clear separation between QIs and SAs. In many data sets it is unclear which attributes are QIs and which are SAs. By separating the sensitive attribute from the QI attributes, Bucketization breaks the attribute correlations between the QIs and the SAs. The anonymized data consist of a set of buckets with permuted sensitive attribute values. Bucketization has been used for anonymizing high-dimensional data.

3.3 Slicing

DATA SLICINGmethod partitions the data both horizontally and vertically, which we discussed previously. The method partitions the data both horizontally and vertically. This reduces the dimensionality of the data and preserves better data utility than bucketization and generalization.

Data slicing method consists of four stages:

- *Partitioning attributes and columns*
  An attribute partition consists of several subsets of A that each attribute belongs to exactly one subset. Consider only one sensitive attribute S one can either consider them separately or consider their joint distribution.
- *Partitioning tuple's and buckets*
  Each tuple belongs to exactly one subset and the subset of tuple's is called a bucket.
- *Generalization of buckets*
  A columngeneralization maps each value to the region in which the value iscontained.
- *Matching the buckets*
  We have to check whether the buckets are matching.

Micro data publishing enable researchers and policy-makers to analyze the data and learn important information.Privacy is a key parameter in sensitive attribute disclosures.For privacy in Micro data publishing generalization and bucketization techniques based on k-anonymity, l-diversity approaches were used.Generalization fails to handle high dimensional data, Bucketization fails to maintain clear separation between quasi-identifying attributes and sensitive attributes.K-anonymity protects against identity disclosures, but it does not provide sufficient protection against attribute disclosures.L-diversity protects against attribute disclosures but fails to prevent probabilistic attacks.So a better system is required that can with stand these failures and offers significant performance rise. For privacy in Micro data publishing a novel technique called slicing is used, which partitions the data both horizontally and vertically.Slicing preserves better data utility than generalization and can be used for membership disclosure protection.Slicing can handle high-dimensional data.For Sliced data to obey the diversity requirement random grouping methods were used.Slicing

algorithm consists of three phases: attribute partitioning, column generalization, and tuple partitioning.Involves the following procedures to attain data anonymity:

a. Attribute Partition and Columns
b. Tuple Partition and Buckets
c. Slicing
d. Column Generalization

These methods compromise on overall data utility to maintain diversity requirement.A better system is required that can that can with stand high-dimensional data handling and sensitive attribute disclosure failures. Fig.1 describes the slicing architecture.

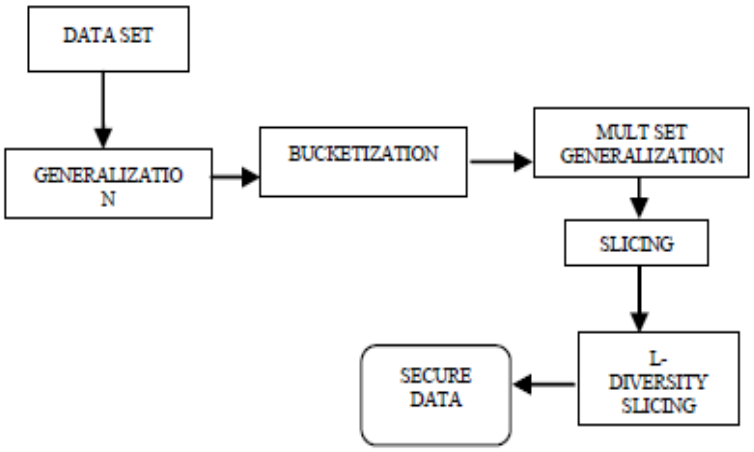

Figure 1: Slicing Architecture

IV. PROPOSED SYSTEM

Data Collection and Data Publishing: A typical scenario of data collection and publishing is described. In the data collection phase the data holder collects data from record owners. As shown in the fig.2 data-publishing phase the data holder releases the collected data to a data miner or the publicwho will then conduct data mining on the published data.

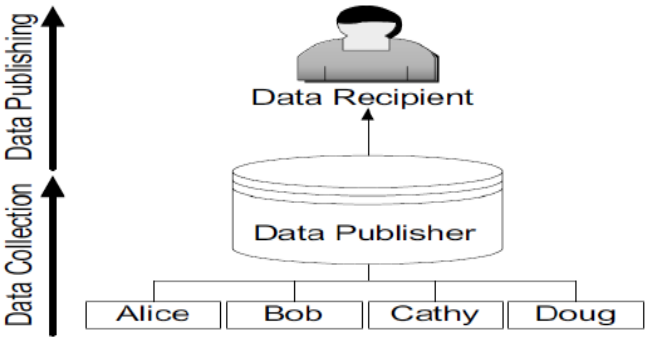

Figure 2: Data collection and Data Publishing

Privacy-Preserving Data Publishing: The privacy-preserving data publishing has the most basic form that data holder has a table of the form: D (Explicit Identifier, Quasi Identifier, Sensitive Attributes, non-Sensitive Attributes) containing information that explicitly identifies record owners. Quasi Identifier is a set of attributes that could potentially identify record owners. Sensitive Attributes consist of sensitive person-specific information. Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories.

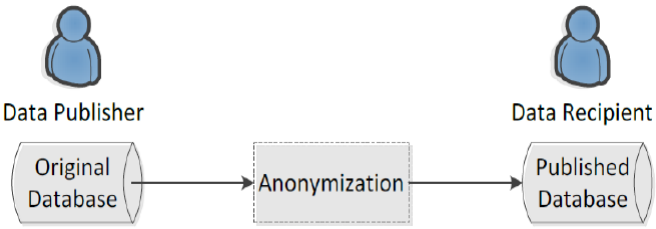

Figure 3: A Simple Model of PPDP

Data Anonymization: Data Anonymization is a technology that converts clear text into a non-human readable form. The technique for privacy-preserving data publishing has received a lot of attention in recent years. Most popular anonymization techniques are Generalization and Bucketization. The main difference between the two-anonymization techniques lies in that bucketization does not generalize the QI attributes.

For privacy in Micro data publishing we stilluse slicing, which partitions the data both horizontally and vertically. Existing Slicing methods compromise on overall data utility to maintain diversity requirement. Therefore, we propose to replacerandom grouping with more effective tuple groupingalgorithms such as Tuple Space Search algorithmbased on hashing techniques. A tuple is defined as avector of ***k*** lengths, where ***k*** is the number of fields ina filter. For example, in a 5-field filter set, the tuple[7, 12, 8, 0, 16] means the length of the source IPaddress prefix is 7, the length of the destination IPaddress prefix is 12, the length of the protocol prefixis 8 (an exact protocol value), the length of thesource port prefix is 0 (wildcard or "don't care"), andthe length of the destination port prefix is 16 (anexact port value). We can partition the filters in afilter set to the different tuple groups. Since the filtersin a same tuple group have the same tuplespecification, they are mutual exclusive and none ofthem overlaps with others in this tuple group. Nowwe can perform the packet classification across allthe tuple has to find the best-matched filter. If multipletuple groups report matches, we resolve thebest-matched filter by comparing their priorities. Thefilters in a tuple can be easily organized into a hashtable, where we use the tuple specification to extractthe proper number of bits from each field as the hashkey. This key can be used for faster indexing, sortingand aprimarily for accurate comparisons. Theefficiency of tuple grouping algorithms enables itsapplication to handle slicing problems that werepreviously prohibitive due to high-dimensional datahandling and sensitive attribute disclosures.

Slicing With Tuple Grouping Algorithm:

The main purpose of Tuple Space Search algorithm is to speed up over all slicing process to support large data.

To replace String based comparisons with Hash based comparisons we used TSS. The hash key can be used for faster indexing,sorting and primarily for accurate comparisons.

TSS increases the significant performance.

The efficiency of TSS algorithm is to handle high dimensional data and sensitive attribute disclosures.

Slicing with Tuple grouping algorithm provides efficient random tuple grouping for micro data publishing. Each column contains sliced bucket (SB) that permutated random values for each partitioned data. The frequency of the value in each one of the scan's-diversity algorithm checks the diversity in each sliced table.

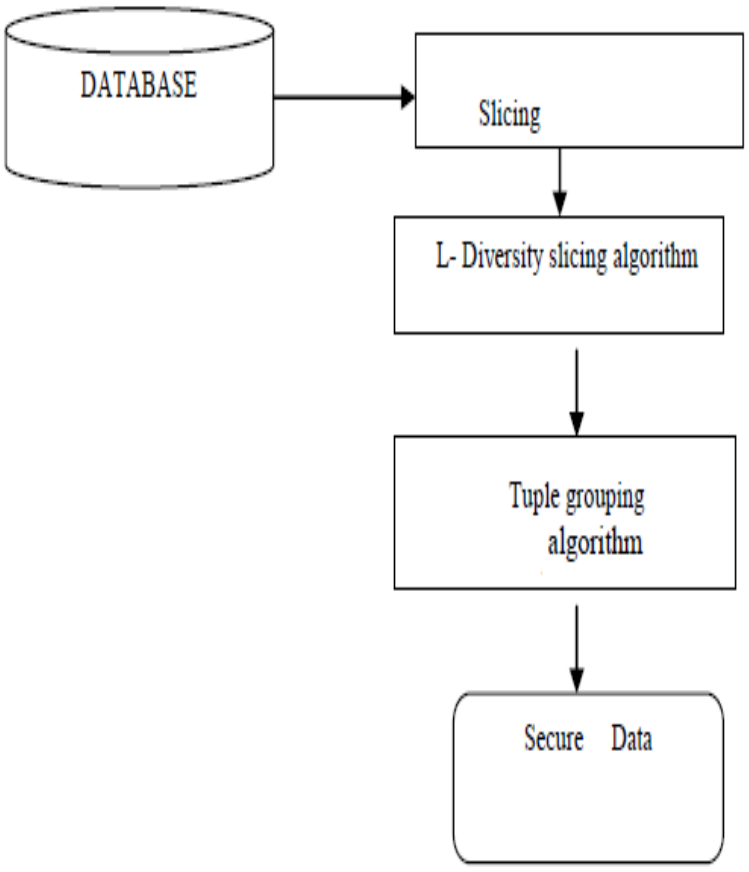

Figure 4. Architecture of slicing with tuple grouping

Fig.5 It describes the functional procedure with respective to the architecture of the slicing with the tuple algorithm.

In Tuple Space for each field one hash key is generated. Using that hash key data is searched.

1.Q=[t], SB=0(NULL).

2.While Q is not empty

3.Remove first Bucket B from Q; Q=Q-B.

4.Split B into two Buckets B1 and B2.

5. for each tuple t in T.

6. for each buckets B in T

Record number of fields =k.

7. partition the tuples depends on the length of fields l(t).

8.construct an access tree.

9. for each fields k inT.

Generate a Hash key(x) using Hash function.

H(x) = x1:q(1), x2:q(2),:::::::::xi:q(i)

10. To find data use Binary search.

11. Find the probability of matching bucket p(t,B) and add it to the list l(t).

12. Q= QU{B1,B2}.

13. else SB=SBU{B}.

14. Return SB.

The algorithm maintains two data structures. 1) a queue of Buckets 2) a set of sliced buckets SB. Initially, Q contains only one bucket which includes all tuples ans SB is empty. Split the bucket into two buckets. For each tuple in T the number of fields in a filter is of length k. Depends on the number of fields partition the tuples. Construct an access tree for these fields. Using Hash generating function generate a hash value to each and every field. The Binary search method is used to find the data is available or not. Compare the generated hash key with the data in access tree. The search process will be done until the item is found or the leaf node exists. If the probability of matched data is found then add it to the list. Then the sliced bucket is constructed.

V. EXPERIMENTS

The original micro data consist of quasi-identifying values and sensitive attributes. As shown in the fig.1 patient data in a hospital.Data consists of Age, Sex, Zip code, disease. A generalized table replaces values.

Age	Sex	Zip code	Disease
22	M	47906	Cancer
22	F	47906	Thyroid
33	F	47905	Thyroid
52	F	47905	Diabetes
54	M	47902	Thyroid
60	M	47902	Cancer
60	F	47904	Cancer

Table.1: Original microdata published.

The recoding that preserves the most information is "local recoding".The first tuple are grouped into buckets and then for each bucketbecause same attribute value may be generalized differently when they appear in different buckets.

Age	Sex	Zip code	Disease
[20-52]	*	4790*	Cancer
[20-52]	*	4790*	Thyroid
[20-52]	*	4790*	Thyroid
[20-52]	*	4790*	Diabetes
[54-64]	*	4790*	Thyroid
[54-64]	*	4790*	Cancer
[54-64]	*	4790*	Cancer

Table.2: Generalized data

Table.2 shows the generalized data of the considered data in the above table. One column contains QI values and the other column contains SA values in bucketization also attributes are partitioned into columns. In the table.3 we describe the bucketization data. One separates the QI and SA values by randomly permuting the SA values in each bucket.

Age	Sex	Zip code	Disease
22	M	47906	Cancer
22	F	47906	Thyroid
33	F	47905	Thyroid
52	F	47905	Diabetes
54	M	47902	Thyroid
60	M	47902	Cancer
60	F	47904	Cancer

Table.3: Bucketized data

The basic idea of slicing is to break the association cross columns, to preserve the association within each column. It reduces the dimensionality of data and preserves better utility. Data slicing can also handle high-dimensional data.

(Age, Sex)	(Zip code, Disease)
(22, M)	(47906, Cancer)
(22, F)	(47906, Thyroid)
(33, F)	(47905, Thyroid)
(52, F)	(47905, Diabetes)
(54, M)	(47902, Thyroid)
(60, M)	(47902, Cancer)
(60, F)	(47902, Cancer)

Table.4: Sliced data

The basic idea of Tuple Space Search algorithm is to produce the results in a less time compared to Slicing.

(Age, Sex)	(Zip code, Disease)
(22, M)	(47906, Thyroid)
(22, F)	(47906, Cancer)
(52, F)	(47905, Diabetes)
(33, F)	(47905, Cancer)
(54, M)	(47902,Diabetes)
(60, M)	(47904, Cancer)
(60, F)	(47902, Cancer)

Table.5: TSS data

In TSS also we are using the same techniques like slicing. But for random generation Slicing cannot able to present data utility and data anonymity in an efficient manner. By using TSS we are presenting better data utility and data anonymity for randomized data.

VI. RESULT ANALAYSIS

To allow direct comparison, we use the l-diversity for two anonymization techniques: slicing and optimized slicing for tuple grouping. We demonstrate experiment demonstrates that:

a. Slicing preserves better data utility than generalization

b. Slicing is more effective than bucketization in workloads involving the sensitive attribute
c. The sliced table can be computed efficiently

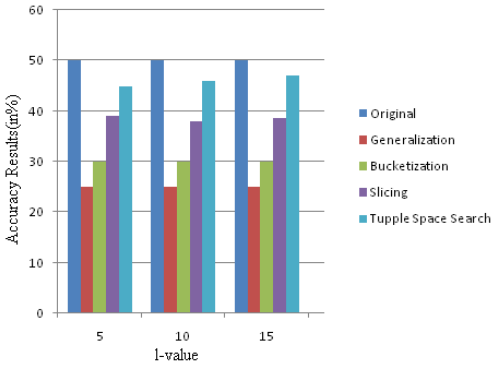

Figure 6: Comparison results of accuracy in datasets.

We compare slicing with optimized slicing in terms of Data utility and Data anonymity.

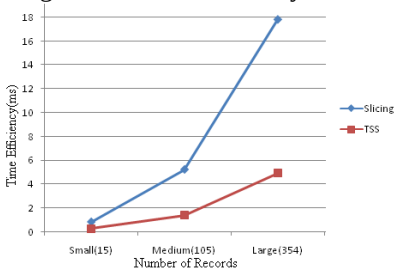

Figure 7: Comparison Results of both slicing and Tuple Space Search in Time efficiency.

In slicing for small data set which consists of 15 records take 0.8ms for processing, whereas in TSS(Tuple Space Search)takes 0.3ms. The time efficiency for medium dataset which consists of 105 records in Slicing is 5.22ms and for TSS takes 1.41ms. The big dataset which contains 354 records is 17.8 and 4.9 to slicing and TSS respectively.

VII. CONCLUSION

Data anonymization is one key aspect of Micro data disclosures as they enable policy-makers to analyze the decision outcomes of issues influencing the business there by influencing the future course of actions. Privacy is a key issue here because inappropriate disclosure of certain data assets will harm the prospects. Prior approaches of data anonymization such as generalization and bucketization (driven by k-anonymity, l-diversity) have been designed for privacy preserving micro data publishing which have several limitations like Generalization's inability to handle high dimensional data and Bucketization failure to maintain clear separation between quasi-identifying attributes and sensitive attributes prompted the development of a novel technique called Slicing, which partitions the data both horizontally and vertically. Although Slicing achieves better data utility and anonymity compared to prior techniques, its sensitive attribute disclosures are based on random grouping, which is not very effective as randomly generating the associations between column values of a bucket significantly lowers data utility. Therefore, we propose to replace random grouping with more effective tuple grouping algorithms such as Tuple Space Search algorithm based on hashing techniques. The computed and obtained sliced data from high dimensional sensitive attributes based on the proposed technique offers significant performance rise. A feasible practical implementation on dynamic data validates our claim.Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. That slicing preserves better data utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. Initially, we consider slicing where each attribute is in exactly one column. Our experiments show that random grouping is not very effective. Proposed grouping algorithm is optimized L-diversity slicing check algorithm obtains the more effective tuple grouping and Provides secure data. Data Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. Another important advantage of slicing is that it can handle high-dimensional data.

REFERENCES

[1] Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy, "Slicing: A New Approach forPrivacy Preserving Data Publishing,"IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, PP:561-574 ,MARCH 2012.

[2] Amar Paul Singh, Ms. DhanshriParihar, " A Review of Privacy Preserving Data Publishing Technique," *International Journal of Emerging Research in Management &Technology*, pp. 32-38, 2013.

[3] M.Alphonsa, V.Anandam, D.Baswaraj, "Methodology of Privacy PreservingData Publishing by Data Slicing," INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND MOBILE APPLICATIONS, pp. 30-34, 2013.

[4] Y. He and J. Naughton, "Anonymization of Set-Valued Data viaTop-Down, Local Generalization," Proc. Int'l Conf. Very Large DataBases (VLDB), pp. 934-945, 2009.

[5] A. Inan, M. Kantarcioglu, and E. Bertino, "Using AnonymizedData for Classification," Proc. IEEE 25th Int'l Conf. Data Eng.(ICDE), pp. 429-440, 2009.

[6] T. Li and N. Li, "On the Tradeoff between Privacy and Utility inDataPublishing," Proc. ACM SIGKDD Int'l Conf. KnowledgeDiscovery and Data Mining (KDD), pp. 517-526, 2009.

[7] T. Li, N. Li, and J. Zhang, "Modeling and Integrating BackgroundKnowledge in Data Anonymization," Proc. IEEE 25th Int'l Conf.Data Eng. (ICDE), pp. 6-17, 2009.

