J. Sc. & Tech. Univ. Peshawar, 2010, 34 (2) 53-58.

INVESTIGATION ON WATER MILL (CHARSI JRANDA) CONSTRUCTED ON WASTEWATER DRAIN (GANDA ERAB) IN RELATION TO POLLUTION MITIGATION

ZAHIDULLAH[1] & MOHAMAD NAFEES[2]

[1]*Department of Environmental Sciences, Alama Iqbal Open University, Islamabad*
[2]*Department of Environmental Sciences, University of Peshawar, Peshawar*

Abstract: The present case-study was an attempt to evaluate the role of watermill in minimizing pollution load in the wastewater drain of Peshawar. The Watermill is situated near Sahabi village: Akbar-Pura and locally it is famous with the name of *Charsi-Jranda*. The watermill was constructed in early 1940s on *Ganda Erab*, a swage drain collecting municipal wastewater from Peshawar City. The drain joins Shahalam River and ultimately the main Kabul River about 2 km downstream of the watermill. The watermill was studied for six months; three months of winter season (low flow) and three months of summer (high flow) season. It was found that *Charsi Jranda* helped to remove 0.78-2.72 kg plastic bags per day, 65.6 to 75.1% of suspended solids and 25.1 to 30.7% BOD_5 per day. *Charsi Jranda* was damaged in 2010 flood and that indigenous free treatment facility was vanished. It is recommended that re-construction of the watermill: *Charsi Jranda* would be environmental friendly; moreover, exploring such kind of other indigenous environment friendly and low cost wastewater treatment facilities are need of the hour.

Key Words: Biological Oxygen Demand (BOD_5), Suspended Solids (SS), Plastic bags, Watermill, Pollution Mitigation

Introduction

Watermill is an old technology usually used for grinding cereal grains. This technology was initiated in ancient times in Egypt, China and Asia. In Asian region, the technology was further upgraded during colonial period (Pujol et al., 2010). Till early 1970s, almost, every village has a watermill. It is still used for grinding purposes in remote areas of Pakistan and India (Sharma et al., 2008). During grinding, it aerates water and increase Dissolved Oxygen level in water up to the maximum which decrease biodegradable organic matter (Alp and Melching, 2011). A wheel is keep turning with the help of a small spillway for which water is stored in the form of a small pond or tank. Here water gets 30 to 60 minutes retention time and enable Suspended Solids (SS) to settle in the form of sludge (Donners, et al., 2002).

Studies revealed that *Ganda Earab* is a big drain contributing a lot of pollution load in Kabul River (Fig. 1). In 1992-93 the Biological Oxygen Demand (BOD_5) of *Ganda Earab* was 220 mg/L during high flow season and 284 mg/L in low flow season (IUCN-DEPM, 1994). In 2004 the BOD level was reached as 335±8.16 mg/L in high flow season and 275±2.45 mg/L during low flow season. The flow recoded at that time was 1.08 – 1.94 M^3/sec (Nafees, 2002). This was enough water to run the watermill and was removing BOD_5 (organic load) through aeration (Apl and Melching, 2011).

The suspended solids recorded in 2004 were ranged 275 to 375 mg/L (Nafees, 2002). Due to heavy suspended load it has got negative effects on Kabul River's flora and fauna (Yousafzai, 2010). The Kabul River water which was used for drinking, bathing and re-creation, now due to heavy municipal and industrial effluents, it is not fit for the aforementioned purposes (Khan and Khan, 1997). In this regards it is one of the theory that Kabul River may be bringing pollution load from Afghanistan (IUCN-DEPM, 1994). Study conducted in 2010 revealed that due to low industrial base, Afghanistan may not be the big contributor in pollution load of Kabul River (Nafees et al, 2010).

To overcome on pollution problems various scientists presented various solutions. One of the solutions is to address pollution sources. For this government of Khyber Pakhtunkhwa constructed four wastewater treatment plants. But unfortunate all of them are not operational and discharging pollution load directly to Kabul River (Nafees, 2002).

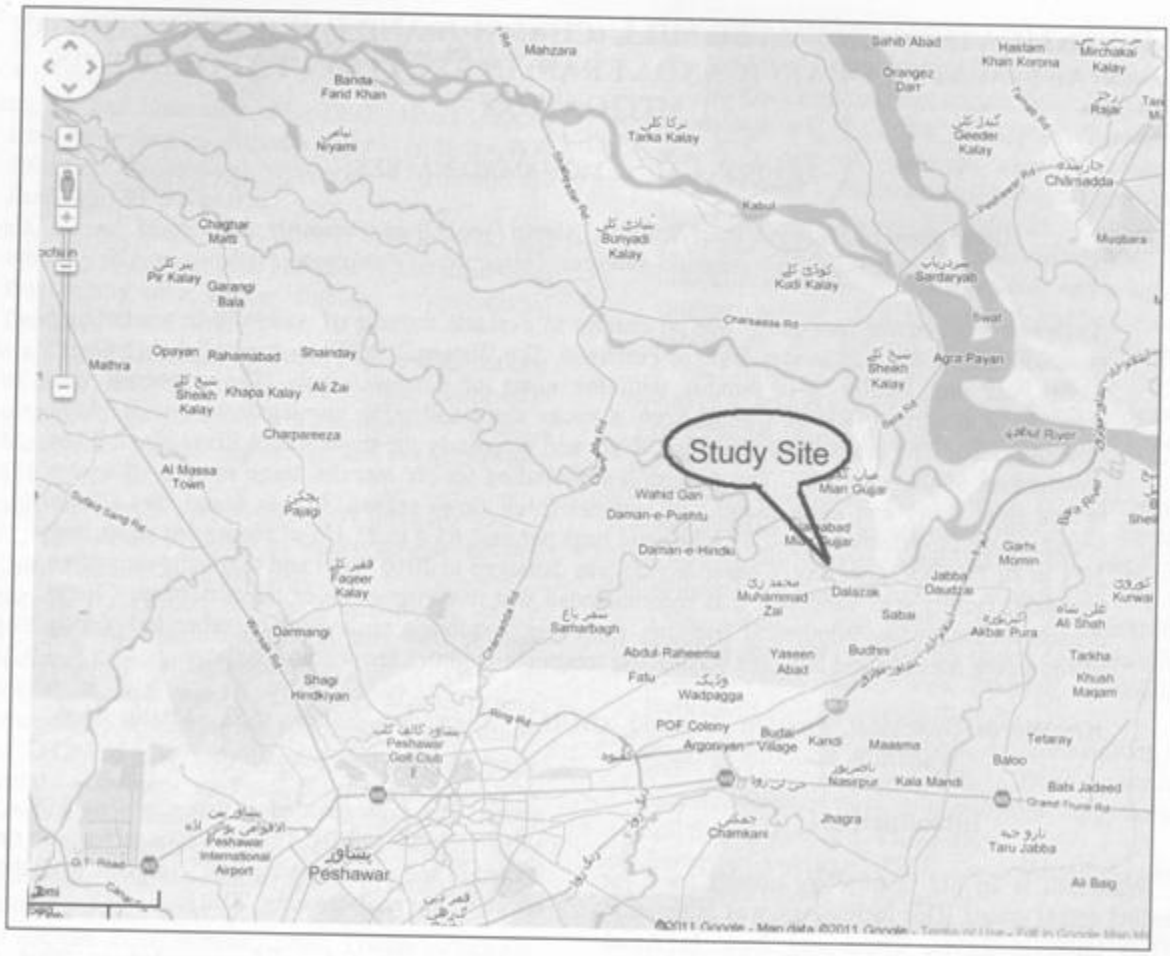

Fig. 1. Map of Study Area

Low cost technology especially of indigenous nature is one of the hopes to reduce pollution load (Helmer, and Hespanhol, 1997). Some countries follow the concept of constructed wetland in which the water is treated in natural way (Gottschalla, et. al, 2007). *Charsi Jranda* is an example which was operated on a stream of wastewater coming from Peshawar.

In this regard, *Charsi Jranda*, designed for grinding purposes, was studied for its role in mitigating pollution load with the objective to re-construct it or make such kind of other simple, local and low cost wastewater treatment facilities to mitigate pollution load in the Kabul River.

Brief on Charsi Jranda

Charsi Watermill (*Jarnda*) was constructed during 1940s. At that time Sir Georg Cunningham was Governor of NWFP now called Khyber Pakhtunkhwa (Douglas, 2009). At that time it was privately owned. The owner was a hard worker and good hunter. He worked as helper with Georg Cunningham and his friends during hunting. One of the friends of Georg Cunningham was engineer. As recognition of his assistance in hunting, he gifted him the design and constructed a water mill on wastewater drain before the confluence point of Shahalam River.

The watermill was working 24 hours a day and was operated by two workers after construction. Each worker was paid 30% of the profit which was equivalent to 20 Kg flour/day or Rs. 500 at present rate. The watermill was grinding at the rate of 800-1000 kg/day (400 – 500 Kg/shift). In this way the mill was saving 50 KW electricity/per day. For maintenance there was another

skilled person. He was supposed to attend the watermill for maintenance once in a month. He was paid to get one day profit. In this way it used to employee four persons. The mill was operated till 2010 and was flooded away in 28 July 2010 flood. During the study period, the role and contribution of *Charsi Jranda* in pollution mitigation has been evaluated, with the objective to think about such low cost indigenous technology to conserve energy and protect environment from further degradation.

Methodology

The watermill was watched and observed once in a week for six months. Three months (June, July and August, 2009) in summer season while three months (January, February and March, 2010) during winter season. In total 24 Water samples (4 samples/ month) were collected and analyzed for physicochemical parameters for comparison with Pakistan National Environmental Quality Standards (Pak-NEQS).

To know about plastic (shopping bags) removal, the plastic bags flowing in the waste water drain were separately collected. The collected plastic bags were then dried and weighted. This exercise was repeated 4 times in a month for six months.

To evaluate removal of Suspended Solids (SS) and Biological Oxygen Demand (BOD_5) 24 samples were collected from the main drain situated 300 meter upstream and downstream of *Charsi Jranda* separately. Wastewater samples were analyzed for SS and BOD_5 by following standard methods for the examination of water and wastewater analysis (Arnold et el., 1992). During the study period, each wastewater sample was a composite of 5 grab samples collected at 2 hours interval. Collected samples were transferred to laboratory on the same day and were analyzed for Suspended Solids. For BOD_5 initial Dissolved Oxygen (DO) was measured after preparing different dilution of 1%, 0.5%, 0.1% and 0.01%. Then after 5 days incubation at 20°C the samples were again analyzed for final DO and thus BOD_5 was calculated.

Results and Discussion

Characteristics of *Ganda Erab*

Ganda Erab consists on municipal wastewater coming from Southeastern part of Peshawar city (Fig. 1). It contributes a lot of pollution load to Kabul River. Among the studied parameters only pH was within permissible range when compared with Pakistan National Environmental Quality Standards (Pak-NEQS). Dissolved Solids were in the range of 923.0±45 to 1647.5±67 mg/L observed in August and January respectively with the average annual rate of 1162.5±45 mg/L. Most of the suspended load consists on organic ranged between 191.6±24 to 391.6±26 mg/L observed in the month of February and July respectively. As organic load consume dissolved oxygen and is a matter of concern. BOD_5 was found in the range of 408.8±38-658.3±43 mg/L. The average annual suspended organic and BOD_5 load was 274.4±22 and 533.5±45 mg/L respectively (Table 1). By comparing the BOD_5 and suspended organic load shows that almost 50% BOD_5 is caused by suspended load while remaining is caused by dissolved solids. In this way a simple sedimentation will remove 50% of BOD_5 along with suspended load.

Table 1. General Characteristics of *Ganda Erab)*

Parameter / Month	pH	Conductivity (uS/cm)	Dissolved Solids (mg/l)	Suspended Solid (mg/l)		BOD (mg/l)
				Organic	Inorganic	
January	7.6±0.2	2353.5±132	1647.5±67	200.4±18	124.8±06	658.3±43
February	7.5±0.2	1779.3±50	1245.5±26	191.6±24	127.3±09	608.5±22
March	7.6±0.2	1604.3±92	1123.0±44	241.7±26	120.3±11	558.7±45
June	7.9±0.3	1504.3±92	1053.0±46	279.3±34	113.2±08	508.2±39
July	8.1±0.1	1404.3±92	983.0±42	341.3±23	96.3±07	458.3±46
August	8.3±0.3	1318.5±80	923.0±45	391.6±26	91.8±09	408.8±38
Average	7.8±0.2	1660.7±90	1162.5±45	274.4±22	112.3±09	533.5±45
Pak NEQS	6.0 - 9.0	-	3500 as total dissolved solids	200 as Total Suspended Solids		80

Plastic Removal

Plastic shopping bags are now integral part of our daily life. These affect our environment in several ways; such as blocking drainage system, and when reached agriculture field cause decrease agriculture production (Njeru, 2006). Plastic shopping bags are dangerous for aquatic life (Lajeunesse, 2004). Therefore, removal of plastic shopping bags from municipal and other waste water drains is very important now days.

Chrsi Jranda was performing this duty free of cost. The pond constructed to regulate water (Fig. 2) and operate the watermill continuously was studied for plastic removal. The average daily plastic removal was 1.66 Kg/day (Table 2). The calculated annual average removal was counted as 605.9 Kg/year. This plastic was removed on regular basis during screening and was helpful to protect Kabul River from plastic load coming in *Ganda Erab.*

Suspended Solids Removal

The minimum annual average removal observed as 100.1 mg/L (348.9-248.8mg/L) while the maximum annual average was 117.3 mg/L (410.2-292.9 mg/L) with the average of 114.1 mg/L (386.0-271.9). In terms of percentage the minimum and maximum removal was 66.2 to 77.6% respectively. The annual average percent removal calculated as 70.8% with a Minimum and maximum annual average of 65.6 to 75.1% respectively (Table 3).

Suspended solids removal was observed as efficient. The suspended solids load of Kabul River is ranged from 255 to 360 mg/L in the low flow season and 600 to 650 mg/L in high flow season. But most of the suspended load of Kabul River main stream was inorganic in nature and have no effect on dissolved oxygen (Nafees, 2002).

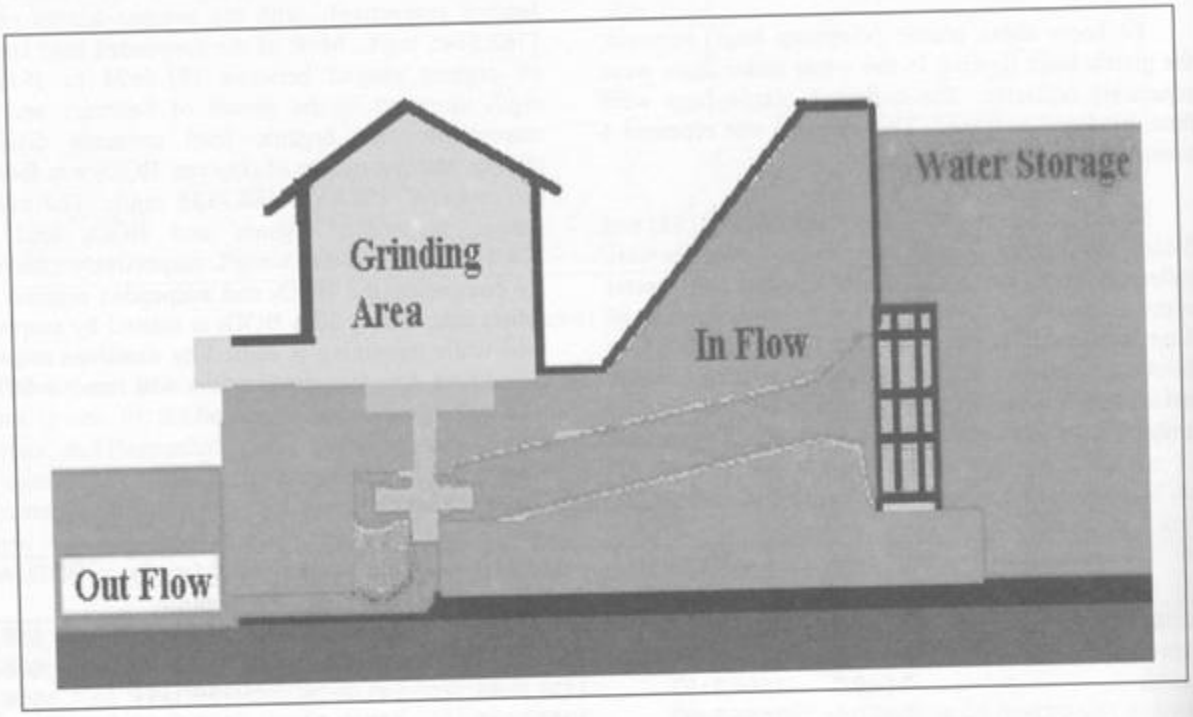

Fig 2. Design Diagram of *Charsi* Watermill

Table 2. Summary of plastic waste removal from *Charsi Jranda* (watermill) during six month time period

Month	Sample	Min	Max	Avg
January	4	0.5	3.5	1.95
February	4	0.6	3.3	2.125
March	4	1.2	3.2	1.775
June	4	0.7	2.1	1.3
July	4	0.8	2.1	1.375
August	4	0.9	2.1	1.425
	Total Samples 24	Average Daily Min. 0.78	Average Daily Max: 2.72	Daily Average 1.66

Table 3. Efficiency of *Charsi Jranda* in Removing Suspended Solids

Month	Up stream of the watermill (mg/L)			Down Stream of the watermill (mg/L)			Percent Removal (%)		
	Min	Max	Ave	Min	Max	Ave	Min	Max	Ave
January	303.0	338.0	325.3	235.9	259.7	252.5	76.7	79.1	77.6
February	272.5	346.0	316.8	195.8	244.0	220.6	65.4	71.9	69.7
March	317.5	391.0	361.8	197.2	266.7	238.2	53.3	77.5	66.2
June	357.5	431.0	390.5	258.5	306.4	284.4	67.4	77.7	73.0
July	399.0	455.0	437.8	280.6	320.6	296.3	62.9	71.1	67.8
August	444.0	500.0	483.5	320.0	359.7	339.6	67.8	73.3	70.3
Average	348.9	410.2	386.0	248.0	292.9	271.9	65.6	75.1	70.8

Table 4. Efficiency of *Charsi Jranda* in Removing BOD_5

	Up stream of the watermill (mg/L)			Down Stream of the watermill (mg/L)			Percent Removal (mg/L)		
	Min	Max	Average	Min	Max	Average	Min	Max	Average
January	612.5	690.6	642.5	175.3	192.5	186.7	25.4	31.1	29.1
Feburary	562.3	640.2	597.4	151.5	190.8	176.2	25.5	32.6	28.7
March	512.5	590.2	547.8	151.6	185.3	161.4	25.4	31.3	29.3
June	462.7	540.5	497.5	117.5	170.3	142.5	24.7	31.5	27.5
July	412.9	490.7	447.6	106.3	145.5	120.3	24.8	29.6	26.6
August	362.3	440.3	397.5	92.8	125.7	104.2	24.7	28.3	25.9
Average	487.5333	565.4167	521.7167	132.5	168.35	148.55	25.08333	30.73333	27.85

Removal of Biological Oxygen Demand

Organic suspended solid and dissolved solids caused BOD_5 (Sonune and Ghate, 2004). During sedimentation, both, organic and inorganic suspended solids are settled down and reduce BOD_5 to a greater extent. Removal of dissolved organic content needs extra aeration (Oller, 2011). In this way, due to the presence of organic suspended solids the BOD_5 removal was termed as efficient. On the average 25.9% to 29.3% with the annual average of 27.85% decrease in BOD_5 was observed (Table 4). The minimum annual load was decrease up to 355 mg/L (487.5-132.5 mg/L). The average maximum removal is calculated as 397.1

mg/L (565.4-168.35). The average annual minimum and maximum decrease is calculated as 25.1 to 30.7% with the overall average daily decrease of 27.85%.

BOD_5 of Kabul River was 2.0 to 5.6 mg/L and 1.0 to 3.7 mg/L during low and high flow seasons respectively (Nafees, 2002). In this way it has got great significance in removing oxidizeable materials (suspended and dissolved solids).

Conclusion

Charsi watermill (*Jranda*) worked efficiently for about 60 years. At the same time, along with the grinding it also served environment by removing basic pollutants like suspended solids, BOD_5 as well as plastic bags from the waste water drain, which consequently reduced the pollution load in Kabul River. It was a wonderful example of indigenous clean technology and was also saving electricity. Therefore, it is strongly recommended to re-erect/reconstruct watermill *Charsi Jranda*,on the same place. Besides other similar low cost, indigenous and environment friendly possibilities should be explored to construct watermills on other main effluent drains before joining Kabul River to minimize/reduce pollution load in the downstream area.

References

Alp, E. and C. S. Melching. 2011. Allocation of supplementary aeration stations in the Chicago waterway system for dissolved oxygen improvement. Journal of Environmental Management, 92(6):1577-1583.

Arnold, E, S. Lenore and D. Andrew. 1992. Standard Methods for the Examination of Water and Wastewater 18 ed, American Public Health Association (APHA), Washington, DC 20005.

Donners, K., M. Waelkens and J. Deckers. 2002. Water mills in the area of Sagalassos: a disappearing ancient technology. Anatolian Studies 52(1):1-17

Douglas B. 2009. Ramparts of Empire: India's North West Frontier and British Imperialism, 1991-1947. Ph.D Thesis, University of Texas at Austin, p-225.

Gottschalla, N., C. Boutinb, A. Crollac, C. Kinsleyc, P. Champagned. 2007. The role of plants in the removal of nutrients at a constructed wetland treating agricultural (dairy) wastewater, Ontario, Canada. Ecological Engineering, 29(2, 1):154-163.

Helmer, R. and I. Hespanhol, eds. 1997. Water Pollution Control: A Guide to the Use of Water Quality Management Principles. Thomson Science and Professional, London. pp-32, 288

International Union for Conservation of Nature and Natural Resources (IUCN) and Department of Environmental Planning and Management (DEPM). 1994. Pollution and the Kabul River an Analysis and Action Plan. A joint Publication of IUCN and DEPM, University of Peshawar, pp-7,84.

Khan, S.A. and M. Khan. 1997. Water Quality Characteristics of the Kabul River in Pakistan Under High Flow Conditions. Journal of Chemical Society Pakistan 19(3):205-210.

Lajeunesse S. 2004. Plastic bags are not created equal because they are meant for different purposes. Journal of Science and Technology, 82(38):51.

Nafees, M. 2002. Environmental Study of Kabul River and its Tributaries in North West Frontier Province (Now Khyber Pakhtunkhwa), Pakistan. M.Phil Thesis, department of environmental Sciences, University of Peshawar, Pp-107-08.

Nafees, M., W. Shah and H. Khan. 2010. Study of Paper Mill for Water Recycling, Hayatabad Industrial Estate, Peshawar. Journal of Science and Technology, University of Peshawar, 34(1): 29-36.

Njeru J. 2006. The urban political ecology of plasticbag waste problem in Nairobi, Kenya. Geoforum, 37(6):1046-1058.

Oller, I., S. Malato, J.A. Sánchez-Pérez. 2011. Combination of Advanced Oxidation Processes and biological treatments for wastewater decontamination—A review. Science of the Total Environment, 409(20):4141-4166.

Pujol T, Jordi Sola, Lino Montoro, Marc Pelegri.2010. Hydraulic performance of an ancient Spanish watermill. Renewable Energy, 35(2): 387–396.

Sharma, R.C. Y. Bisht, R. Sharma and D. Singh.2008. Gharats (watermills): Indigenous device for sustainable development of renewable hydro-energy in Uttrakhand Himalayas. Renewable Energy 33 (10):2199–2206.

Sonune, A. and R. Ghate. 2004. Developments in wastewater treatment methods. Desalination 167:55–63.

Yousafzai, A.M., A. R. Khan and A.R. Shakoor. 2008. An Assessment of Chemical Pollution in River Kabul and its Possible Impacts on Fisheries. Pakistan Journal of Zoology, 40(3):199-210.

J. Sc. & Tech. Univ. Peshawar, 2010, 34 (2) 59-66.

ACHIEVING K-ANONYMITY WITHOUT GENERALIZATION AND SUPPRESSION

AZHAR RAUF, SAEED MAHFOOZ, SHAH KHUSRO & SHABIR AHMAD

Department of Computer Science, University of Peshawar, Peshawar, Pakistan

Abstract: Publishing data for analysis purposes and maintaining the privacy of individuals in the published data is the need of the day. One commonly used approach is k-anonymity method. K-anonymity is normally achieved either by generalizing or suppressing the data due to which data utility is compromised. If data is more generalized, less utility of data is achieved and vice versa. In this paper we propose a novel approach which is based on pivot columns in which two tuples having same attribute values are merged into one tuple and then published. Our mechanism not only provides protection of individual's privacy, but also preserves data utility because no generalization and suppression are involved. Our algorithm also satisfies the properties of p-sensitivity, l- diversity and t-closeness. Experiments confirm that our novel approach allows significant and effective data analysis over the conventional method of generalization and suppression.

Keywords: Pivot Attribute, Anatomy, Privacy, Quasi Identifiers, Microdata

Introduction

In this modern age, the information sharing is very important for finding the public health and demographic information of the people of an area, country or region. There are many organizations that publish data in un-aggregated form which is also known as microdata (Xiaokui & Yufei, 2006). This data is in the form of table and each record which is also known as row in database terminology represents the information of an individual. This data is required to statistics department and other organizations that are interested to analyze the data and take useful information about that particular area, for research purposes or for the allocation of public funds. However, private information of an individual such as salary, medical condition, and location should not be uniquely identified from the microdata. Traditionally at the time of releasing microdata, an individual's unique information such as name, address, contact number, and social security number are removed from microdata to protect his private information. However, this approach also does not guarantee the privacy of an individual in the released microdata. According to recent estimation, in the United States 87% population can be identified uniquely with the help of seemingly innocent attributes like date of birth, gender, and 5-digit zip-code (Sweeney, 2002). Consider Table 1 which seemingly shows protected data of individual's medical record of a hospital. However, if we join this de-identified table with publically available databases like census data or voter registration lists, with the help of zip-code, age and sex generally known as quasi identifiers, one can identify an individual from the table. The unique identification of the medical record of the Massachusetts's governor with the help of quasi identifiers like gender, zip-code, date of birth and diagnosis is one famous linking attack example. The governor's quasi identifiers were linked with Massachusetts voter registration records which include name, zip-code, gender and dated of birth (Sweeney, 2002). The Massachusetts's governor William Weld, medical record was stored in the Group Insurance Commission (GIC) data (March, 19, 19971997 #27). When his medical record was linked with the Cambridge Voter list as he was living in Cambridge, six people date of birth was same in which three were men; and he was the only one in his 5-digit Zip-code. Thus unique identification of the Governor Weld's medical record was made possible by linking attack.

Table I. Microdata

Job	Birth	Sex	Zip-code	Disease
Clerk	1975	M	4350	HIV
Manager	1955	M	4350	Flu
Clerk	1955	F	5432	Flu
Worker	1955	F	5432	Fever
Worker	1975	M	5543	Flu
Technician	1940	M	5543	Fever

Traditionally k-anonymity is achieved by generalization and suppression. This research work proposes a novel approach to achieve k-anonymity without generalization and suppression. The level of distribution is less as compared to the generalization and suppression approach, which results improved information disclosure while maintaining same level of privacy. In our approach two or more tuples are combined into a single tuple. The data is merged in a tuple on the basis of common attribute value among these tuples. We named the common attribute as pivot attribute. Our novel approach reduces the level of confidence of the intruder to find individual's information in released data. We suggest an algorithm that satisfies k-anonymity without generalization and suppression. We show that results of our approach are better than the existing approaches.

The rest of paper is organized as follows. In the following section, we discuss the existing work about achieving k-anonymity via generalization and suppression. We show the limitation of generalization and suppression techniques. We discuss the anatomy approach, a new method of privacy preserving. Then we explain some basic definition and notations used in our proposed approach. Later on we discuss the proposed model, algorithm and then present the experimental setup and its results. We finally discuss comparison of our work with existing techniques and conclude our paper.

Related Work

Linking attacks using quasi-identifiers is a well known privacy problem. Researchers adopted different approaches and methodologies in order to overcome these attacks. Smarati and Sweeny proposed a definition of privacy called *k-anonymity* (Sweeney, 2002), (Samarati, 2001). A k-anonymous table satisfies the property that each record in the table is similar and it is not possible to distinguish it from at least k-1 other records with respect to every set of quasi identifier attributes. This means at least k-records share these values in such a way that an individual cannot be uniquely identified by linking attacks. Table II shows a 2-anonymous view corresponding to Table I. Here the "Disease" is a sensitive attribute which is maintained without any change in this example.

Table II. 2-Sensitive 2-Anonymous

Job	Age	Sex	Zip-code	Disease
White-collar	*	M	4350	HIV
White-collar	*	M	4350	Fever
*	1955	F	5432	Flu
*	1955	F	5432	Fever
Blue-collar	*	M	5543	Flu
Blue-collar	*	M	5543	Fever

In recent years, k-anonymity is widely discussed by various researchers as a solution for individual privacy in released data and a number of algorithms have been proposed for implementing it. The main approach used by these researchers is generalization and suppression. Samarati (Samarati, 2001) came up with an algorithm that uses a binary search on the domain generalization hierarchy to find minimal k-anonymous table. Bayardo and Agrawal (Bayardo & Agrawal, 2005) presented an algorithm that converts a fully generalized table and specializes it to the minimal k-anonymous table. LeFevre et al. (Kristen & Ramakrishnan, 2005) came up with a priori computation and bottom up technique and proposed an algorithm that achieves k-anonymity with minimum generalization. A. Machanavajjhala et al. (Ashwin et al., 2007) pointed out two attacks, the homogeneity and background knowledge attacks that can still reveal individuals' sensitive information due to lack of diversity in sensitive attribute. X.Hu et al. (Xinping Hu, 2009) suggested that instead of complete generalization and suppression of data, only those tuples should be generalized which belong to sensitive attributes hence, reducing information loss and increasing data utility.

All approaches proposed by different researchers so for achieve k-anonymity via generalization or suppression (Sweeney, 2002), (Pin & Chen, 2010; Samarati, 2001; Bayardo & Agrawal, 2005; Charu, 2005; Kristen & Ramakrishnan, 2005; Xiaokui & Yufei, 2006; Ashwin, Daniel et al. 2007; Ninghui & Venkatasubramanian, 2007; Deng & Xiao-Jun, 2008; Sun & Ping, 2008; Talouki & Baraani, 2009; Xinping Hu, 2009; Junwei Zhang & Yuan, 2010; Ren & Yang, 2010; Ninghui Li, June, 2010). In all the techniques it is assumed that each attribute has its own conceptual generalization hierarchy or taxonomy tree. In this hierarchy lower level provides more detailed

information then higher level in the same type of hierarchy. For example, Zip-code 5575 is at lower level in hierarchy and provides more details then 557* which is one step higher in hierarchy.

Comparing Table I with Table II it is obvious that generalization and suppression though provides privacy, looses the information utility. As the basic purpose of publishing data is to use it for analysis and research purposes, generalization and suppression may affect the quality of reporting because of information loss.

A. From generalization and Suppression to Anatomy Approach

A new approach called anatomy approach is proposed (Xiaoxun Sun, Hua Wang et al., 2009) instead of generalizing or suppressing a data set and then publishing it. In this approach it is suggested that if two tables with join attributes are published then the join between these two tables will be lossy and due to lossy join the private information will be concealed. According to this technique, the original microdata table is divided into two tables. One table consists of non sensitive attributes along with group id as shown in Table III. The second table consists of sensitive attribute along with group id as shown in Table IV. When these two tables are combined with the help of group id then due to lossy join they will result the Table V. It is difficult for attacker to know that who is the actual HIV patient, because the manager and clerk are corresponding to same bag of disease {HIV, Flu}.

B. Defects in the Anatomy Approach

Although the Anatomy approach has given new directions to privacy preserving in publish data, but it has several shortcomings. As the basic purpose of publishing microdata is to use it for research purposes and to gain useful information from this published data. However, if the query against published data does not retrieve correct information then the published data is not suitable for analysis. The Anatomy approach, though provides an individuals' privacy but on the cost of providing incorrect information in some cases. In this approach instead of publishing a single table two tables are published joined by group id attribute.

Table III. Non Sensitive (NSS) Table

Job	Birth	Sex	Zip-code	Group ID
Clerk	1975	M	4350	1
Manager	1955	M	4350	1
Clerk	1955	F	5432	2
Worker	1955	F	5432	2
Worker	1975	M	4350	3

Table IV. Sensitive (SS) Table

Disease	Group ID
HIV	1
Flu	1
Flu	2
Fever	2
Flu	3
Fever	3

Table V. Resulting Join Table

Job	Birth	Sex	Zip-code	Disease
manager	1955	M	4350	HIV
clerk	1975	M	4350	HIV
manager	1955	M	4350	Flu
clerk	1975	M	4350	Flu
worker	1955	F	5432	Flu
clerk	1955	F	5432	flu
worker	1955	F	5432	Fever
clerk	1955	F	5432	Fever
technician	1940	M	5543	Flu
worker	1975	M	5543	Flu
technician	1940	M	5543	Fever
worker	1975	M	5543	fever

Following are some of the scenarios showing the issues related to the Anatomy approaches:

Scenario No 1

Suppose in a particular region if the number of HIV cases is required, then following query against TABLE III and TABLE IV is posed.

```
SELECT nss.job, nss.birth, nss.sex, nss.zip_code,
ss.disease
FROM nss, ss
WHERE nss.group_id=ss.group_id;
```

Note: NSS refers to Table III in which Quasi Identifier and group id attributes are present and SS refers to Table IV in which Sensitive attribute and group id attributes are present.

The result of above query is TABLE V in which the number of HIV cases are two, which shows wrong results because in original microdata the HIV case is only one as shown in Table I.

Scenario No 2

Similarly, if the number of flu patients in Zip-code 4350 is required, then following query against TABLE III and TABLE IV is posed:

```
SELECT nss.job, nss.birth, nss.sex, nss.zip_code,
ss.disease
FROM nss, ss
WHERE zip_code= 4350 and disease ='Flu';
```

The result of above query is TABLE VI, which has six records. This is not correct as the actual cases of flu patients in Zip-code 4350 are two in original micro data as shown in Table I.

Table VI. Resulting Join Table

Job	Birth	Sex	Zip-code	Disease
Clerk	1977	M	4350	Flu
Clerk	1953	M	4350	Flu
Clerk	1977	M	4350	Flu
Manager	1945	M	4350	Flu
Manager	1977	M	4350	Flu
Manager	1953	M	4350	Flu

Scenario No 3

Suppose it is required to know the number of fever patients whose birth date is 1955, the following query is posed against Table III and Table IV.

```
SELECT nss.job, nss.birth, nss.sex, nss.zip_code,
ss.disease
FROM nss, ss
WHERE birth= 1955 and disease ='Fever';
```

The result of above query is TABLE VII which has six records. This is also not correct information because the actual cases of fever patients whose birth date is 1955 is only one in original microdata as shown in Table I.

Table VII. Resulting Join Table

Job	Birth	Sex	Zip-code	Disease
Manager	1955	M	4350	Fever
Clerk	1955	F	5432	Fever
Worker	1955	F	5432	Fever
Manager	1955	M	4350	Fever
Clerk	1955	F	5432	Fever
Worker	1955	F	5432	Fever

The above scenarios prove that Anatomy approach has some serious issues in term of publishing microdata as it provides incorrect information.

Basic Definitions

Suppose 'T' is a table and a subset of some large population and in this table each tuple represents an individual of the population. This table is of the form T $\{Q_i, SA, N_i\}$. For Example T is a medical dataset. Here Q_i represents quasi-identifiers (such as birth, sex, zipcode), SA is sensitive attribute and N_i denotes other attribute of 'T'.

Quasi-Identifier Attribute Set (Qi): It is a set of non sensitive attributes except identifier attributes that is used in linking attack to find an individual from a dataset. Examples are set of attribute like {gender, age, zip-code, date of birth, nationality} which can be used in linking attacks to uniquely identify an individual.

Equivalence class: It is a set of tuples with respect to Qi set in which each tuple is indistinguishable from rest of the tuple in the group.

Sensitive Attributes (SA): Attributes that contain sensitive information and known to everyone but need to be protected. Examples of sensitive attributes are medical condition, salary and credit card number which are used for analysis purposes.

Pivot Attributes (P_i): It is an attribute inside Quasi-Identifiers i.e. $P_i \in Qi$ and on the basis of which two tuples are merged into one tuple. The pivot attribute may be zip-code, age, birth date. It depends on data set and Qi group that which attribute should be selected as a Pivot Attribute.

Property of K-anonymity: A table T is said to having the k-anonymity property with respect to quasi

identifier set Qi if every tuple of T with respect to Qi is greater than or equal to k.

Proposed Technique and Algorithm

In this section we present an algorithm called Attribute Selection for Merging (ASM). The algorithm merges two or more tuples having similar attributes in microdata into one tuple. As different tuples are merged into one tuple so there are fields which hold multiple values in one tuple. Those values which occur in one field of tuple are separated with the help of comma.

The algorithm is shown in figure 1. The input of this algorithm is original microdata table 'T' and two pivot attribute 'PP' and SP as input parameters. The PP stands for primary Pivot which merges the different tuples on the basis of same Qi value. SP stands for Secondary pivot which holds the property if in sensitive value there is a lack of diversity then that value must be merged with other tuple on the basis of this secondary pivot. The output of this algorithm is K-anonymous table **T'**. These pivots attributes PP and SP are selected among Qi identifiers depending on the situation and data set.

A. **Working of Algorithm:** First the algorithm searches the whole data set on the basis of Primary Pivot (PP) and those tuples having the same value on the basis of this pivot attribute is merged in to one tuple. The sensitive attribute values are checked if they lack diversity then these tuples are merged in another tuple on the basis of Secondary pivot. As one field of tuple will contain more than one value these values are separated with the help of comma as shown in Table VIII.

As no generalization is involved in our novel approach due to which the precision of publishing table is improved.

Algorithm: ASM

Input: a microdata T, pivot Attributs PP & SP

Output: publishing table T1 satisfying k-anonymity

1. if $|T| \leq 1$
2. return
3. for each ti and tj in T if ti[PP] = tj[PP]
4. concate tj to ti on the basis of PP and place it in T1 //T1 is table that would be published
5. if SA of ti lack diversity then
6. concate t_j to t_k on the basis of SP
7. if t[QI] in T1 is greater than one values separate it by comma.
8. T1 is our table for publishing

Fig. 1. ASM.

Table VIII. Resulting pivot Colum Table

Job	Birth	Sex	Zip-Code	Disease
Clerk, Manager	1975,1955	M	4350	HIV,Flu
Clerk, worker	1955	F	5432	Flu,Fever
Worker,Technician	1975,1940	M	5543	Flu,Fever

Experiment Setup and Results

The experiment is conducted on the Adult Database from the UCI Machine Learning Repository (Frank, 2010). The database contains 45,222 tuples and it is about US Census data. Records with missing values are removed and 30,000 tuples are selected for experiments. The Adult database contains 15 attributes and 7 attributes out of them are selected for experiments. The 'Occupation' is considered as a sensitive attribute. We used an implementation of our proposed algorithm ASM which is a novel algorithm that does not generalize or suppress data. The software used in experiments includes Microsoft Windows XP, Oracle Database 10g and PL/SQL. All experiments are run on a 3.0 GH processor with 2GB Main Memory. The database used in our experiments is described in fig.2.

	Attribute	Distinct Values
1	Age	74
2	Work Class	7
3	Education	16
4	Marital Status	7
5	Native country	41
6	Sex	2
7	Occupation	14

Fig. 2. Description of Adult dataset

Here are some of the experimental results of our proposed technique. We selected Age as our Primary Pivot [PP] and Native country as our Secondary Attribute SA. The size of the equivalence class was 5 so we got 6000 group. Each group was satisfying the individual's anonymity. Those tuples of equivalence classes which was exploiting the l-diversity property, was shuffled to satisfy the anonymity property. The resulting table obtained was in the form of Table VIII. For experiment purposes to check the correctness of data and to see whether the data is in anonymous form in our proposed technique we posed some queries against original microdata used in our experiment and then the same queries were posed against the published data converted in anonymized form by our algorithm. For example to see the occupation of a person whose age is 40 years, education bachelor and native country is United States we posed the following query against published data converted by our technique:

Select age, education, native_country, occupation
From adult
Where age = 40 AND education ='bachelor' And native_country ='United States';

The result of this query was {craft repair, prof speciality, admin, clerical, tech support} so it was not possible for an intruder to guess that which one is the actual occupation of a person based upon the quasi identifiers age, education, native_country. Because the information was matched to the same bag of values that is {craft repair, prof speciality, admin, clerical, tech support}. In the proposed technique the l-diversity in each equivalence class is also maintained which means that data is also protected from homogeneity and background knowledge attacks.

As in the proposed technique there is no generalization and suppression involved so the information loss is zero percent and data utility is hundred percent.

Comparison with Other Techniques

Lyengar (Iyengar, 2002) showed that genetic algorithm can be used to attack on highly flexible and highly combinatorial formulation of k-anonymity. As this algorithm is an incomplete stochastic search method it does not guarantee on solution quality. The proposed method of Sweeny (Sweeney, 2002) also known as the datafly approach is greedy approach. This approach generates the list of frequencies and those combinations which are less than occurrences are iteratively generalized. Iterative greedy approach of Sweeny like incomplete stochastic approach, does not guarantee quality solution. Samarati (Samarati, 2001) and Sweeney (Sweeney, 2002) both have proposed algorithms for k-Anonymization. To identify the optimal generalization Sweeney's algorithm exhaustively examines all potential generalizations to achieve minimal anonymity. This approach is impractical even on ordinary sized datasets. The Samarati's algorithm identifies all "k-minimal" generalizations and among these minimal generalizations according to specific criteria optimal k-Anonymization exists. As this algorithm makes use of monotonicity property and binary search property on the generalization lattice which avoids searching the entire generalization space exhaustively. The main problem in this approach is that the number of k-minimal generalizations becomes too

large to count efficiently as these generalizations become exponential.

(Winkler, 2002) has used the algorithm of simulated annealing for attacking the problem but it also does not provide the guarantee of its efficacy. As our proposed algorithm is using the simple linear search technique first for primary Attribute PP and then in case of lack of diversity it uses the secondary pivot to group the data. The summary of different known techniques in term of practicality and solution guarantees is given in the Figure 3.

Technique	Practicality	Solution Guarantee
Sweeny MinGen	Not Practical	Yes
Sweeny Datafly	Practical	Not
Samarati All MinGen	Not Practical	Yes
Wrinkler Anneal	Possible	Not
Lynger GA	Yes	Not
Our Proposal	Practical	Yes

Fig 3. Comparison of Different Approaches

Conclusion and Future Work

P-sensitive k-anonymity via anatomy is a novel approach that satisfies individual privacy without generalization and suppression. However as shown in this paper, the major disadvantage of this approach is delivery of wrong information to analyst. In this paper we proposed a novel method called achieving k-anonymity without generalization and suppression. Our extensive experimental results show that our proposed model delivers better results in terms of accuracy and data utility.

References

Ashwin, M., K. Daniel, et al. (2007). "L-diversity: Privacy beyond k-anonymity." ACM Trans. Knowl. Discov. Data 1(1): 3.

Bayardo, R. J. and R. Agrawal (2005). "Data Privacy through Optimal k-Anonymization." Proc of the 21st International Conference on Data Engineering, IEEE Computer Society. pp.217-228.

Charu, C. A. (2005). "On K-anonymity and the curse of dimensionality." Proceedings of the 31st international conference on Very large data bases. Trondheim, Norway, VLDB Endowment: 901-909.

Deng, J.-J., Ye, and Xiao-Jun (2008). "Algorithm for multidimensional k-anonymity by R tree." Jisuanji Gongcheng / Computer Engineering 34(1): 80-82.

Frank, A. & Asuncion, A. (2010) UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Adult.]. Irvine, CA: University of California, School of Information and Computer Science.

Iyengar, V., S. (2002). "Transforming data to satisfy privacy constraints." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. Edmonton, Alberta, Canada, ACM: 279-288.

Junwei Zhang, J. Y., Jiapei Zhang, and Y. Yuan (2010). "KIDS:K-anonymization data stream base on sliding window." IEEE 2nd International Conference on Future Computer and Communication (ICFCC) Wuhan, China. 2: V2-311 - V2-316.

Kristen, L., David J. DeWitt, Raghu, and Ramakrishnan (2005). "Incognito: efficient full-domain K-anonymity." Proc of ACM SIGMOD International conference on Management of data, Baltimore, Maryland. pp 49-60.

Ninghui, L., Tiancheng L., and S. Venkatasubramanian (2007). "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity." IEEE 23rd International Conference on Data Engineering (ICDE). Istanbul: 106-115.

Ninghui Li, T. L., Suresh Venkatasubramanian (June, 2010). "Closeness: A New Privacy Measure for Data Publishing." IEEE Transactions on Knowledge and Data Engineering 22(7): 943-956.

Pin, L., Yu, W., and N. Chen GSSK: "A Generalization Step Safe Algorithm in Anonymizing Data." IEEE International Conference on Communications and Mobile Computing (CMC), 2010. 1: 183-187.

Ren, X. and J. Yang (2010). "Research on Privacy Protection Based on K-Anonymity." International Conference on Biomedical Engineering and Computer Science (ICBECS). 1-5.

Samarati, P. (2001). "Protecting Respondents' Identities in Microdata Release." IEEE Trans. on Knowl. and Data Eng. 13(6): 1010-1027.

Sun, X. W., Hua Li, Jiuyong Truta, Traian Marius Li and Ping (2008). "(p+, α)-sensitive K-anonymity: A new enhanced privacy protection model." 8th IEEE International Conference on Computer and Information Technology. Sydney, Australia: 59-64.

Sweeney, L. (2002). "K-anonymity: a model for protecting privacy." Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10(5): 557-570.

Sweeney, L. (2002). "Achieving k-anonymity privacy protection using generalization and suppression." Int'l Journal on Uncertainty, Fuzziness, and Knowledge-Base Systems 10(5): 571-588.

Talouki, M. A. N., M.-a. and A. Baraani (2009). "K-anonymity privacy protection using ontology." In proc. IEEE 14th International CSI Computer Conference (CSICC'09): 682-685.

W. E. Winkler. "Using Simulated Annealing for k-anonymity." Research Report Series (Statistics #2002-7), U. S. C. B., 2002.

Xiaokui, X. and T. Yufei (2006). "Personalized privacy preservation." Proceedings of the 2006 ACM SIGMOD international conference on Management of data. Chicago, IL, USA, ACM: 229-240.

Xiaoxun Sun, Hua Wang, et al. (2009. p 199-205). "Achieving P-Sensitive K-Anonymity via Anatomy." IEEE International conference on e-Business Engineering.

Xinping Hu, Z. S., Yingjie Wu, Wenyu Hu, Jiancheng Dong (2009). "K-Anonymity Based on Sensitive Tuples." First International Workshop on Database Technology and Applications: 91-94.