DATA PRESERVATION AND BREACH DETECTION ON SOCIAL NETWORK

Temitope B. Salau[1], **Manoj K. Gupta**[2], **Abdullahi A. Haruna**[3]

[1] Department of Computer Science and Engineering, School of Engineering and Technology, Sharda University,
Greater Noida, Uttar Pradesh 201306, India.

[2] Department of Computer Science and Engineering, Krishna Institute of Engineering & Technology, Ghaziabad,
Uttar Pradesh 201206, India.

[3] Department of Computer Science and Engineering, School of Engineering and Technology, Sharda University,
Greater Noida, Uttar Pradesh 201306, India.

Abstract

The increasing recognition of social networks has brought about a productive research area in information withdrawal and data mining. Statistics from various fields such as security organizations, government organizations, etc. reflect that the amount or rate at which sensitive data is exposed have grown at the double in recent years over social network such as MySpace, FaceBook, Twitter etc. This exposure could be caused by, human mistakes which is the main cause of loss of data. It is possible that organizations and individuals are alerted of any data exposure caused by human mistakes with the existence of an adequate solution. There exists a requirement of effective anonymization techniques to solve these problems. This paper exploited the methodology referred to as KDLD anonymity i.e., K-degree-l-diversity model which is applied to an extracted FaceBook data (with the aid of a social network data analyzing tool, Gephi). The anonymization tool, ARX which already has the features and working of KDLD is used to anonymize the data extracted. The anonymized data was then transferred to NodeXL for further and clearer representation.

Keywords: *Anonymization, k-degree-l-diversity, Privacy-Preserving Data Mining, Privacy-Preserving Data Publishing, Social Network Data*

1. Introduction

Social Networks such as MySpace, Facebook etc. have become gradually more accepted applications in the Web. Lots of users are registered on them; here every user is linked to a bunch of other users through friendship, professional relationship (members of the same organizations), etc. There is a large amount of personal information available on the Web about any individual that is generated either by an individual or by others. Recently, social network data have been made available to the public in some kind of way. Hereby making Preserving privacy in publishing social network data become an important [1] Nowadays, more Web users are generating information that is meant to be private. An organization or an individual could use the generated information to make assumptions or conclusions about a person. The social interaction and performances between users involved on social network can be viewed as a graph whereby each vertex represents a user, and the social associations and actions are represented by the edges [2]. The results of the structure of the graph have lots of vertices (i.e. users or social actors) also lots of edges (i.e. social associations). Previously many methods and techniques were proposed to preserve privacy of social network. The most efficient is privacy-preserving data publishing which mainly deals with anonymizing graph data. In the view of solving the above problems associated with the anonymization of social network data, some methods and techniques were proposed, which include the following: Attack Model [3] and [4], Privacy Model [5] and Data Utility [6]. In this paper the re-identification risk model was combined with k-degree-l-diversity model to ensure the privacy of a Facebook data. Section 2 discusses the literature review of similar works, proposed methodology and previous works. Section 3 captures the implementation of this work, parameters and tools used. Section 4 showcases the various results and outcomes from the implementation of this work while discussion, conclusion and future work are elaborated in section 5.

2. Related Work

2.1 Attack Model

Given the anonymized network data, foes for the most part depend on foundation information to de-anonymize people and learn connections between de-anonymized people. Six sorts of the foundation learning were recognized by [7], which are: qualities of vertices, vertex degrees, join relationship, neighborhoods, installed sub-diagrams and chart measurements. An algorithm called Seed-and-Grow was proposed by [3] to distinguish clients from an anonymized social diagram, construct exclusively with respect to chart structure. The calculation first distinguishes a seed sub-diagram which is either planted by an attacker or disclosed by plot of a little group of clients, and afterward develops the seed bigger in view of the enemy's current information of clients' social relations. Various types of attack model include:

2.1.1 Mutual Friendship Attack: this method depends on the amount of common friends of two associated people on a social network. In a social network website, for example, Facebook or MySpace, an attacker can secure the amount of friends of an individual [8]. Besides, the foe can likewise remove the friendship connection between two people from the interaction data freely accessible on the site. Accordingly, utilizing the vertex degrees of two people and their fellowship connection, the enemy can issue a friendship attack on the social network that was published to re-distinguish the vertices comparing to an individual and his companion and related vertex data, for example, hobbies, exercises and religious convictions [2]. The Fig. 1 below illustrates an example of mutual friendship attack model.

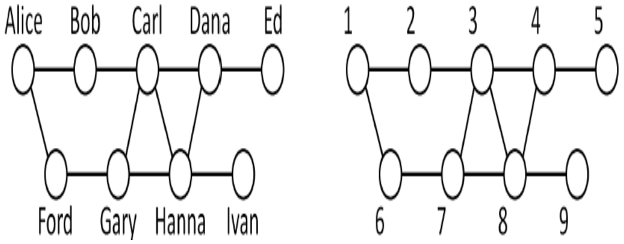


Fig. 1a Original social network Fig. 1b anonymized social network

2.1.2 Degree Attack: here, every person in a social network is obliged to interact with more than one vertex identification as well as a group personality, and the group personality mirrors some sensitive information about the person. It has been demonstrated that, taking into account some foundation learning about vertex degree, regardless of the fact that the attacker can't unequivocally distinguish the vertex relating to an individual, group data and neighbourhood data can at present be derived [9]. Fig. 2 represents and illustrates the degree attack model.

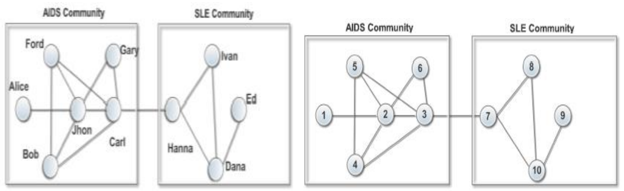


Fig. 2a Original network data Fig. 2b naïve Anonymized network

2.2 Privacy Model

This depends on the great k-anonymity model, various security models have been proposed for graph data. A portion of the models have been condensed in the overview made by [10], for example, k-degree, k-neighbourhood, k-automorphism, k-isomorphism, and k-symmetry. In order to shield the security of relationship from the mutual friend attack, a variation of k-anonymity, called k-NMF secrecy was presented by [5]. NMF is a property characterized for the edge in an undirected basic graph, which represents the amount of mutual friends between the two people connected by the edge. In the event that a system fulfills k-NMF anonymity then for every edge e, there will be in any event k-1 different edges with the same number of mutual friends as e. It can be ensured that the probability of an edge being recognized is not more prominent than 1/k. Fig. 3 depicts the privacy model technique (K-NMF anonymity).

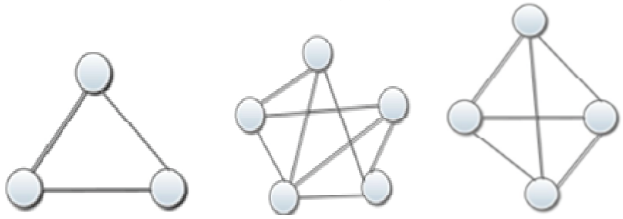


Fig. 3a 3-NMF Fig. 3b 4-NMF Fig.3c 6-NMF

2.3 Data Utility Model

With respect to network data anonymization, the implication of data utility is: whether and to what extent properties of the graph are preserved. [11] summarize three types of properties considered in current studies, which are

2.3.1 Graph Topological Properties**:** which are characterized for applications that aim at the analysis of the properties of the graph. Different measures have been proposed to demonstrate the structure qualities of the network.

2.3.2 Graph Spectral Properties: The range of a graph is normally characterized as the set of eigen-values of the graph's adjacency matrix or other inferred matrices, which has close relations with numerous graph properties.

2.3.3 Aggregate Network Queries: An aggregate network query calculates the aggregate on some paths or subgraphs satisfying some query conditions. The accuracy of answering aggregate network queries can be considered as the measure of utility preservation.

2.4 Privacy-Preserving Data Publishing

There is a large amount of personal information available on the Web about any individuals that is generated either by them or by others. Furthermore, more users on the

Web are generating information that is deemed private. A corporation or a person can use that information to make decisions about a particular individual [12]. PPDP in the context of social networks mainly deals with anonymizing graph data. In the view of solving the above problems associated with the anonymization of social network data, some methods and techniques were proposed

2.4.1 *K*-Anonymity: *k*-anonymization algorithms for network data publishing perform edge insertion and/or deletion operations, and they try to reduce the utility loss by minimizing the changes on the graph degree sequence. [13] Considered that the degree sequence only captures limited structural properties of the graph and the derived anonymization methods may cause large utility loss. They propose utility loss measurements built on the community-based graph models, including both the flat community model and the hierarchical community model, to better capture the impact of anonymization on network topology.

2.5 Data Breach Detection Technique

Identity Disclosure (ID) is a critical problem in private data publishing, and has been widely studied in previous work. Recently, the ID problem is attracting increasing attention in social network analysis [14]. Intrusion Detection Systems (IDSs) are widely deployed to defend against large-scale attacks and security threats. Existing IDSs, however, often trigger too many alerts daily [15]. Moreover, the false alerts are mixed up with the true ones [16], making it more difficult to identify the real hidden attacks.

According to a report from Risk Based Security (RBS) [17], the number of leaked sensitive data records has increased dramatically during the last few years [18]. Detecting and preventing data leaks or breach requires a set of complementary solutions, which may include data-leak detection, data confinement, stealthy malware detection, and policy enforcement [17]. Some of the techniques used to detect data breach include:

2.5.1 Fuzzy Fingerprint Method and Protocol: Here, the DLD provider obtains digests of sensitive data from the data owner. The data owner uses a sliding window and Rabin fingerprint algorithm [19] to generate short and hard to- reverse (i.e., one-way) digests through the fast polynomial modulus operation. The sliding window generates small fragments of the processed data (sensitive data or network traffic), which preserves the local features of the data and provides the noise tolerance property. Rabin fingerprints are computed as polynomial modulus operations, and can be implemented with fast XOR, shift, and table look-up operations. The Rabin fingerprint algorithm has a unique min-wise independence property [20], which supports fast random fingerprints selection (in uniform distribution) for partial fingerprints disclosure. This whole phenomenon is referred to as *Shingles and Fingerprints.*

2.5.2 Frequent pattern mining technique: This technique is highly suitable for alert analysis with the ability of deriving frequent item-sets and association rules. Especially, algorithm FP-Tree is good at efficient mining of frequent patterns in large database, which applying a compressed, frequent pattern tree FP-Tree structure for mining frequent patterns without candidate generation. The FP-Tree algorithm mainly includes the following two algorithms: FP-Tree construction and FP-Growth [21]. Emphasis will be put on method of mining association rules among protected alert’s attributes for retrenching and filtering intrusion events.

2.5.3 Deep Packet Inspection Technique (DPI): Network data-leak detection (DLD) typically performs deep packet inspection (DPI) and searches for any occurrences of sensitive data patterns. DPI is a technique to analyze payloads of IP/TCP packets for inspecting application layer data, e.g., HTTP header/content. Alerts are triggered when the amount of sensitive data found in traffic passes a threshold. The detection system can be deployed on a router or integrated into existing network intrusion detection systems (NIDS). Straightforward realizations of data-leak detection require the plaintext sensitive data. However, this requirement is undesirable, as it may threaten the confidentiality of the sensitive information. If a detection system is compromised, then it may expose the plaintext sensitive data (in memory). In addition, the data owner may need to outsource the data-leak detection to providers, but may be unwilling to reveal the plaintext sensitive data to them. Therefore, one needs new data-leak detection solutions that allow the providers to scan content for leaks without learning the sensitive information [17].

3. Proposed Method

The model that is proposed to be used for the achievement of this work is *k*-degree-l-diversity (KDLD) anonymity. The Proposed approach defines the k-degree-l-diversity anonymity model that considers the protection of structural information as well as sensitive labels of individuals.

3.1 *K-degree-l-diversity* (KDLD) anonymity

In this method, the goal is to prevent a user from re-identifying other users and finding out sensitive or delicate information about them. *K*-degree-*l*-diversity (KDLD) model is used for securely issuing a labelled graph, and then develop corresponding graph anonymization algorithms with the least distortion to the properties of the original graph, such as degrees and distances between nodes [5]. *K*-degree-*l*-diversity (KDLD) model will produce anonymization methodology based on adding noise nodes. A new algorithm is produced by adding noise nodes into the original graph with the consideration of introducing the least distortion to graph properties [22]. In addition to KDLD, re-identification risk was combined with it on the ARX anonymization tool in order to reduce the re-identification of users on the social network. In ***K-degree-l-diversity* (KDLD) anonymity,** the social network is modelled as a graph G = (V, E, L, $L_{V,}$ L_E), where V represents a set of vertices, E ⊆ V x V represents a set of edges, L represents a set of labels and the labelling function is describe as L_V: V→ L which assigns each vertex a label, a labelling function L_E: E→L which assigns each edge a label. For a graph G, V (G), E (G), L (G), L_V (G) and L_E (G), these are the set of vertices, edges, labels, vertex labelling function in G and the edge labelling function in G respectively.

3.2 Proposed Architecture

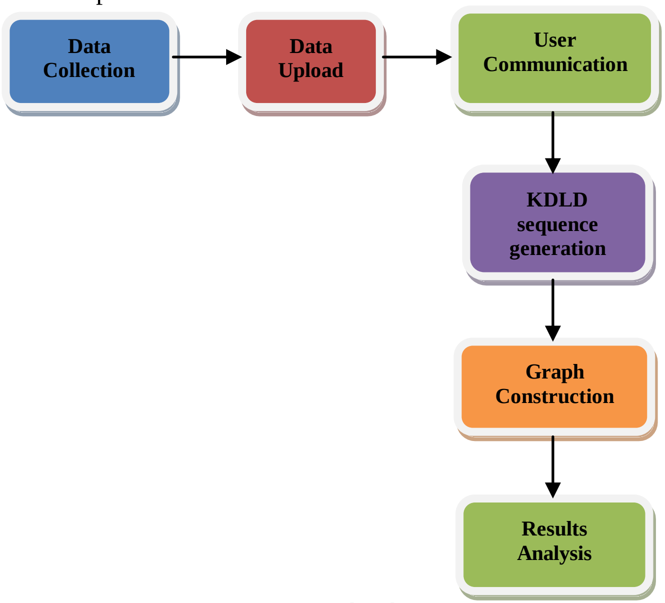


Fig. 4 Proposed Architecture

3.2.1 Proposed Architecture Implementation Process
The implementation process includes the following steps:

- Collection of data from the source: this is the process whereby the data collector collects data from the data users in order for the data to be published. The data was collected from a social media network, precisely facebook.
- Data Validation: this is done by the data miner (i.e. data cleaning: this is done with Statistica software) in order to confirm or ensure that the data is clean, correct and useful for publishing.
- Data uploading: here the data uploaded onto gephi tool for further analysis of the graph and for a better structure.
- User Communication: in this process the node and edges of the graph were analyzed, i.e. different individuals involved on the network and their activities, also with whom they commune with and in which society.
- KDLD Anonymization: after further analysis and coversion of the data from .gdf file to excel file, ARX tool which has an in-built anonymization models was exploited for the anonymization of the data.
- Graph Construction: after anonymization, the graph of the anonymized network was constructed with use of ARX tool or NodeXL

  Results Analysis**:** this displays the outcome of the models application such as data utility, information loss, identification risk, etc.

4. Experiment Implementation

The tool used to download the data from the internet is called ***Gephi:*** which helps in the better exploration and understanding of the data graph for better analysis and better graph structure. It downloads the data in form of .gdf file extension. ***NodeXL*** was used to convert the .gdf file of the data downloaded from the internet into and excel file in order for the data to be properly anonymized on the anonymization tool. The anonymization tool used for the anonymization of the data is ***ARX anonymization tool***. In this implementation, the ARX data anonymization tool implements a novel globally-optimal anonymization algorithm, known as Flash algorithm, which constructs a search space and determines the transformation with minimal information loss. The Flash algorithm traverses the generalization lattice in a bottom-up breadth-first manner and constantly generates paths which branch like lightning flashes. The ARX anonymization tool has an in-built of different privacy models such as *k*-anonymity, *l*-diversity, *t*-closeness, *k*-Map, (e, d)-differential privacy, δ-presence, δ-disclosure privacy etc. The data was anonymized using ***KDLD*** i.e. *k*-degree-*l*-diversity, *k*-degree 2 and 2-*l*-diversity also average re-identification risk was included to the privacy models in order to reduce the risk of individuals' identification on the network.

4.1 Flash Algorithm

Flash algorithm is implemented in ARX anonymization tool. It iterates over all levels in the lattice, starting at level *0*. It enumerates all transformations on each level and calls findPath(node) if a transformation has not been tagged already. This method implements a greedy depth-first search towards the top node. The search terminates when either the top node is reached or the current node does not have a successor that is not already tagged.

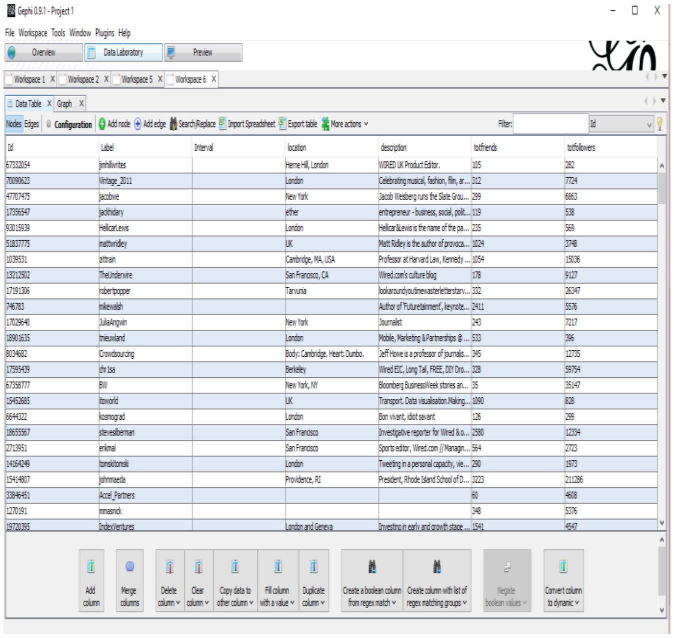
```
PriorityQueue pqueue = new PriorityQueue();
for (int i = 0; i < lattice.height; i++) {
  for (Node node : sort(level[i])) {
    if (!node.isTagged()) {
      pqueue.add(node);
      while (!pqueue.isEmpty()) {
        Node head = pqueue.poll();
        if (!head.isTagged()) {
          check(findPath(head), pqueue);
        }
      }
    }
  }
}

private void findPath(Node node){
  List path = new List();
  while (path.head() != node){
    path.add(node);
    for(Node up : sort(node.getSuccessors())){
      if (!up.tagged){
        node = up;
        break;
      }
    }
  }
  return path;
}

private void check(List path, PriorityQueue pqueue) {
  int low = 0; int high = path.size() - 1;
  while (low <= high) {
    int mid = (low + high) / 2;
    Node node = path.get(mid);
    if (!node.isTagged()) {
      checkAndTag(node);
      if (!node.isAnonymous()) {
        for (final Node up : node.getSuccessors()) {
          if (!up.isTagged()) {
            pqueue.add(up);
          }
        }
      }
    }
    if (node.isAnonymous()) high = mid - 1;
    else low = mid + 1;
  }
}
```

4.2 Experiment Analysis

Fig. 5 illustrates the data downloaded from the web and how it was uploaded on Gephi tool in form .gdf format. After uploading the social network data, the tool analyses the data and represents it in form of graph with nodes and edges in Fig. 6. Each cluster of nodes with distinguished color represents users that belong to a particular community or society. Fig. 7a depicts the lattice graph representation of the social network data after it has been anonymized with KDLD method in addition with Re-identification risk model with the ARX anonymization tool, from that result it can be seen that there was minimal information loss. In Fig. 7b the data that was anonymized was then further analyzed with the Gephi tool, from the graph representation, it can be seen that each user cannot be identified or be linked to a particular society because they are all interwoven. Fig 8 shows that the re-identification risk of each record in the social network data after anonymization reduced, the re-identification risks were very high before anonymization, this can be seen on the left-hand partition of Fig. 8 while the right-hand partitions depicts the re-identification risk after anonymization. The exported and published data are shown in Fig 9 and Fig. 10 respectively.



Id	Label	Interval	location	description	totfriends	totfollowers
67332054	jimhillwrites		Herne Hill, London	WIRED UK Product Editor.	105	282
70090623	Vintage_2011		London	Celebrating musical, fashion, film, ar...	312	7724
47707475	jacobwe		New York	Jacob Weisberg runs the Slate Grou...	299	6863
17356547	jackhidary		ether	entrepreneur - business, social, polit...	119	538
93015939	HelicarLewis		London	Helicar&Lewis is the name of the pa...	235	569
51837775	mattwridley		UK	Matt Ridley is the author of provoca...	1024	3748
1039531	zittrain		Cambridge, MA, USA	Professor at Harvard Law, Kennedy ...	1054	15036
13212502	TheUnderwire		San Francisco, CA	Wired.com's culture blog	178	9127
17191306	robertpopper		Tarvunia	lookaroundyoutimewasterletterstarv...	332	26347
746783	mikewalsh			Author of 'Futuretainment', keynote...	2411	5576
17029640	JuliaAngwin		New York	Journalist	243	7217
18901635	tnieuwland		London	Mobile, Marketing & Partnerships @ ...	533	396
8034682	Crowdsourcing		Body: Cambridge. Heart: Dumbo.	Jeff Howe is a professor of journalis...	345	12735
17595439	chr1sa		Berkeley	Wired EIC, Long Tail, FREE, DIY Dro...	328	59754
67358777	BW		New York, NY	Bloomberg BusinessWeek stories an...	35	35147
15452685	itoworld		UK	Transport. Data visualisation.Making...	1090	828
6644322	kosmograd		London	Bon vivant, idiot savant	126	299
18655567	stevesilberman		San Francisco	Investigative reporter for Wired & o...	2580	12334
2713951	erikmal		San Francisco	Sports editor, Wired.com // Managin...	564	2723
14164249	tomskitomski		London	Tweeting in a personal capacity, vie...	290	1973
15414807	johnmaeda		Providence, RI	President, Rhode Island School of D...	3223	211286
33846451	Accel_Partners				60	4608
1270191	mmasnick				348	5376
19720395	IndexVentures		London and Geneva	Investing in early and growth stage ...	1541	4547

Fig. 5 Social Network Data Uploaded on Gephi in .gdf format

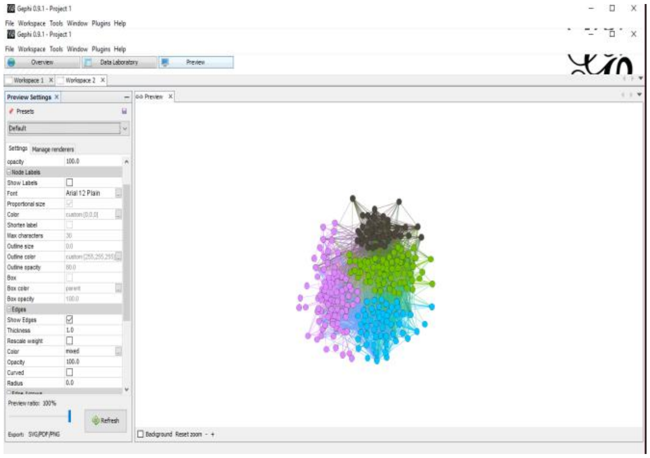
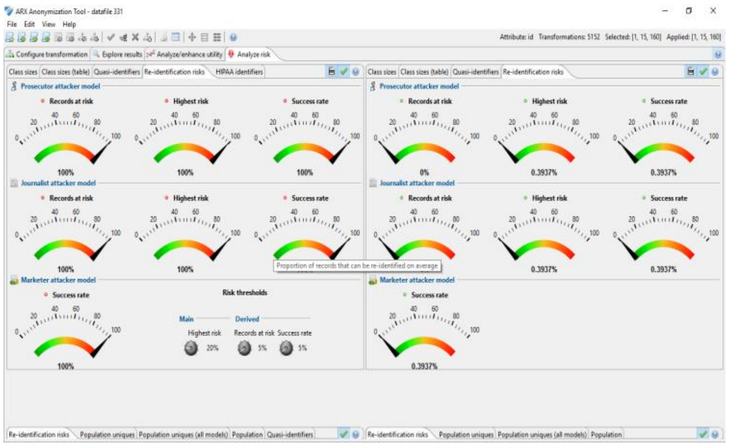


Fig. 6 Social network data graph before Anonymization

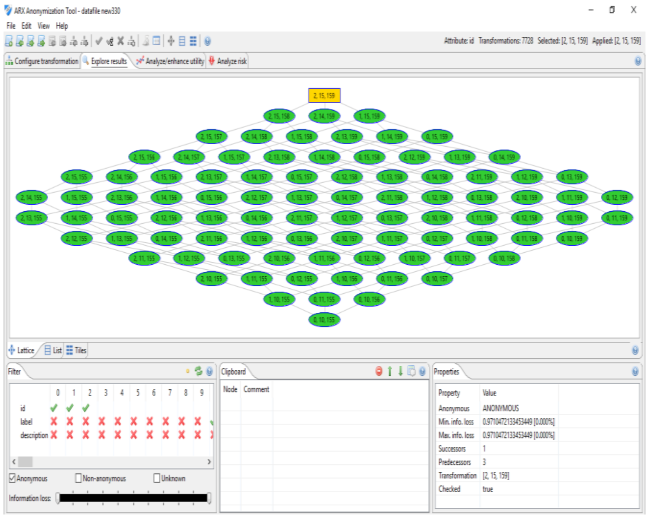


Attribute: id Transformations: 5152 Selected: [1, 15, 160] Applied: [1, 15, 160]

	Records at risk	Highest risk	Success rate
Prosecutor attacker model	100%	100%	100%
Journalist attacker model	100%	100%	100%
Marketer attacker model			100%
Prosecutor attacker model	0%	0.3937%	0.3937%
Journalist attacker model	0.3937%	0.3937%	0.3937%
Marketer attacker model			0.3937%

Risk thresholds — Main: Highest risk 20%; Derived: Records at risk 5%, Success rate 5%

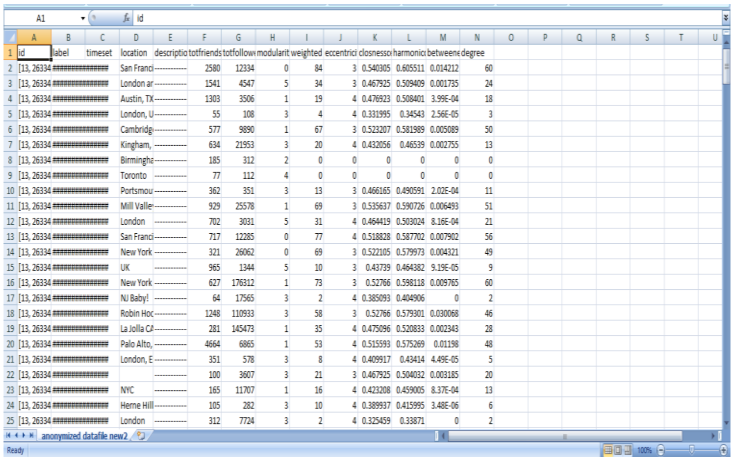
. 8 Social Network Data Re-Identification Before and After Anonymization



Attribute: id Transformations: 7728 Selected: [2, 15, 159] Applied: [2, 15, 159]

Property	Value
Anonymous	ANONYMOUS
Min. info. loss	0.9710472133453449 [0.000%]
Max. info. loss	0.9710472133453449 [0.000%]
Successors	1
Predecessors	3
Transformation	[2, 15, 159]
Checked	true

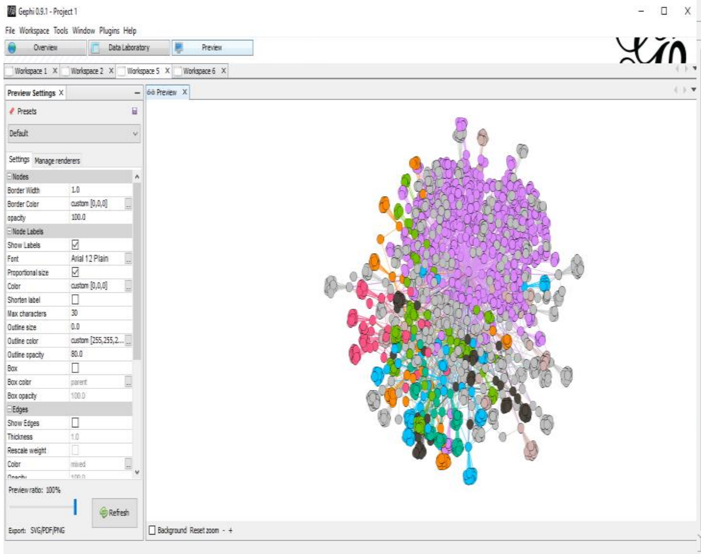
Fig. 7a Social Network Data Graph after Anonymization


A1 id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id	label	timeset	location	descriptio	totfriends	totfollow	modularit	weighted	eccentrici	closnesso	harmonic	betweene	degree
2	[13, 26334	###############		San Franci	-----------	2580	12334	0	84	3	0.540305	0.605511	0.014212	60
3	[13, 26334	###############		London ar	-----------	1541	4547	5	34	3	0.467925	0.509409	0.001735	24
4	[13, 26334	###############		Austin, TX	-----------	1303	3506	1	19	4	0.476923	0.508401	3.99E-04	18
5	[13, 26334	###############		London, U	-----------	55	108	3	4	4	0.331995	0.34543	2.56E-05	3
6	[13, 26334	###############		Cambridg	-----------	577	9890	1	67	3	0.523207	0.581989	0.005089	50
7	[13, 26334	###############		Kingham,	-----------	634	21953	3	20	4	0.432056	0.46539	0.002755	13
8	[13, 26334	###############		Birmingha	-----------	185	312	2	0	0	0	0	0	0
9	[13, 26334	###############		Toronto	-----------	77	112	4	0	0	0	0	0	0
10	[13, 26334	###############		Portsmou	-----------	362	351	3	13	3	0.466165	0.490591	2.02E-04	11
11	[13, 26334	###############		Mill Valle	-----------	929	25578	1	69	3	0.535637	0.590726	0.006493	51
12	[13, 26334	###############		London	-----------	702	3031	5	31	4	0.464419	0.503024	8.16E-04	21
13	[13, 26334	###############		San Franci	-----------	717	12285	0	77	4	0.518828	0.587702	0.007902	56
14	[13, 26334	###############		New York	-----------	321	26062	0	69	3	0.522105	0.579973	0.004321	49
15	[13, 26334	###############		UK	-----------	965	1344	5	10	3	0.43739	0.464382	9.19E-05	9
16	[13, 26334	###############		New York	-----------	627	176312	1	73	3	0.52766	0.598118	0.009765	60
17	[13, 26334	###############		NJ Baby!	-----------	64	17565	3	2	4	0.385093	0.404906	0	2
18	[13, 26334	###############		Robin Hoc	-----------	1248	110933	3	58	3	0.52766	0.579301	0.030068	46
19	[13, 26334	###############		La Jolla CA	-----------	281	145473	1	35	4	0.475096	0.520833	0.002343	28
20	[13, 26334	###############		Palo Alto,	-----------	4664	6865	1	53	4	0.515593	0.575269	0.01198	48
21	[13, 26334	###############		London, E	-----------	351	578	3	8	4	0.409917	0.43414	4.49E-05	5
22	[13, 26334	###############			-----------	100	3607	3	21	3	0.467925	0.504032	0.003185	20
23	[13, 26334	###############		NYC	-----------	165	11707	1	16	4	0.423208	0.459005	8.37E-04	13
24	[13, 26334	###############		Herne Hill	-----------	105	282	3	10	4	0.389937	0.415995	3.48E-06	6
25	[13, 26334	###############		London	-----------	312	7724	3	2	4	0.325459	0.33871	0	2

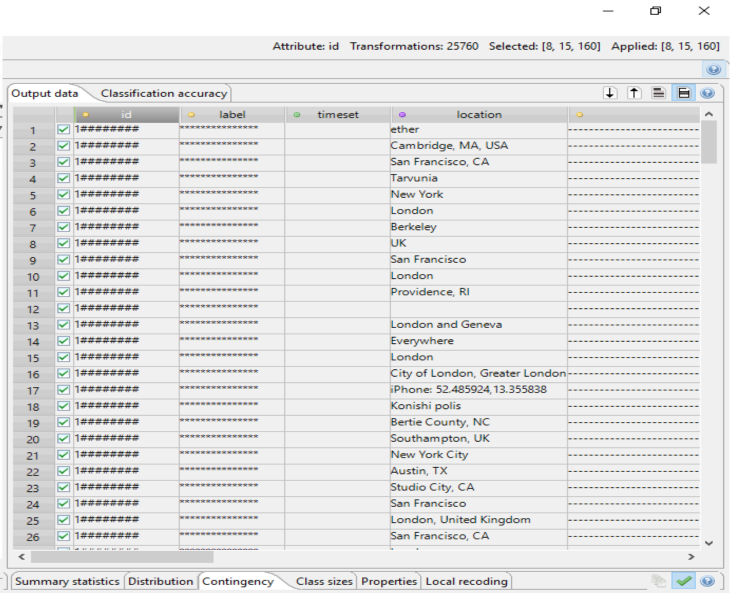
anonymized datafile new2
Ready

Fig. 9 Exported data after Anonymization



Nodes: Border Width 1.0; Border Color custom [0,0,0]; opacity 100.0
Node Labels: Show Labels ☑; Font Arial 12 Plain; Proportional size ☑; Color custom [0,0,0]; Shorten label □; Max characters 30; Outline size 0.0; Outline color custom [255,255,2...; Outline opacity 80.0; Box □; Box color parent; Box opacity 100.0
Edges: Show Edges □; Thickness 1.0; Rescale weight □; Color mixed
Preview ratio: 100%

Fig. 7b Social Network Data graph on Gephi after Anonymization



Attribute: id Transformations: 25760 Selected: [8, 15, 160] Applied: [8, 15, 160]

		id	label	timeset	location	
1	☑	1########	***************		ether	------------------------
2	☑	1########	***************		Cambridge, MA, USA	------------------------
3	☑	1########	***************		San Francisco, CA	------------------------
4	☑	1########	***************		Tarvunia	------------------------
5	☑	1########	***************		New York	------------------------
6	☑	1########	***************		London	------------------------
7	☑	1########	***************		Berkeley	------------------------
8	☑	1########	***************		UK	------------------------
9	☑	1########	***************		San Francisco	------------------------
10	☑	1########	***************		London	------------------------
11	☑	1########	***************		Providence, RI	------------------------
12	☑	1########	***************			------------------------
13	☑	1########	***************		London and Geneva	------------------------
14	☑	1########	***************		Everywhere	------------------------
15	☑	1########	***************		London	------------------------
16	☑	1########	***************		City of London, Greater London	------------------------
17	☑	1########	***************		iPhone: 52.485924,13.355838	------------------------
18	☑	1########	***************		Konishi polis	------------------------
19	☑	1########	***************		Bertie County, NC	------------------------
20	☑	1########	***************		Southampton, UK	------------------------
21	☑	1########	***************		New York City	------------------------
22	☑	1########	***************		Austin, TX	------------------------
23	☑	1########	***************		Studio City, CA	------------------------
24	☑	1########	***************		San Francisco	------------------------
25	☑	1########	***************		London, United Kingdom	------------------------
26	☑	1########	***************		San Francisco, CA	------------------------

Fig. 10 Published data after Anonymization

5. Conclusion

Privacy is now an important issue in data mining; PPDM is an important aspect of data mining that aims to provide security for secret information from unauthorized disclosure. Privacy- preserving data Mining is important because nowadays threat to privacy is becoming serious. Some data mining techniques can be used to predict highly sensitive knowledge or instance from a given huge quantity of data. The proposed a model will not only prevent the revelation of users' identity but also the disclosure of sensitive information in users' profiles. In this k-degree-l-diversity model for privacy preserving social network data publishing, the main difference between previous and this system is that it mainly focuses on noise node adding algorithm to construct a new graph from the original graph with the constraint of introducing fewer distortion to the original graph. After the implementation of the proposed model it was observed that there was minimal information loss hereby bringing about maximal data utility. Also the percentage of individuals that were exposed to high re-identification risk of disclosure reduced drastically, and the individuals affected by lower re-identification risk of disclosure increased due to the addition of the model of re-identification risk.

5.1 Future Work

In future, a technique is hoped to be developed that flags an alert whenever social network data is breached or whenever an unauthorized access is made to the data early enough or on-time and also to reduce the quantity of information loss during anonymization hereby increasing data utility in order for a more adequate and efficient preservation. Also, protocols should be designed to help data publishers publish a unified data together to guarantee privacy.

References

[1] Zhou, B & Pei, J 2010, "The k-anonymity and l-diversity approaches for privacy preservation in social network against neighborhood attack", *Springer.*

[2] C.-H. Tai, P. S. Yu, D.-N. Yang, and M.-S. Chen,``Privacy preserving social network publication against friendship attacks,'' in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1262-1270.

[3] W. Peng, F. Li, X. Zou, and J. Wu, ``A two-stage deanonymization attack against anonymized social networks,'' *IEEE Trans. Comput.*, vol, 63, no. 2, pp. 290-303, 2014.

[4] T. Zhu, S. Wang, X. Li, Z. Zhou, and R. Zhang, `` Structural attack to anonymous graph of social networks,'' *Math. Problems Eng.*, vol. 2013, Oct. 2013, Art. ID 237024.

[5] C. Sun, P. S. Yu, X. Kong, and Y. Fu. 2013, ``Privacy preserving social network publication against mutual friend attacks.'' [Online]. Available: http://arxiv.org/abs/1401.3201

[6] N. Medforth and K. Wang, ``Privacy risk in graph stream publishing for social network data,'' in *Proc. IEEE 11th Int. Conf. Data Mining (ICDM)*, Dec. 2011, pp. 437-446.

[7] B. Zhou, J. Pei, and W. Luk, ``A brief survey on anonymization techniques for privacy preserving publishing of social network data,'' *ACM SIGKDD Explorations Newslett.*, vol. 10, no. 2, pp. 12_22, 2008.

[8] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography," in *Proc. 16th Int. Conf.World Wide Web (WWW'07)*, New York, NY, USA, 2007, pp. 181–190, ACM.

[9] Xu, L, Jiang, C, Wang, J, Yuan, J & Ren, Y 2014, 'Information Security in Big Data: Privacy and Data Mining', *IEEE Translations Access*, vol. 2, no. 10, pp. 1149-1176.

[10] M. I. Hafez Ninggal and J. Abawajy, ``Attack vector analysis and privacypreserving social network data publishing,'' in *Proc. IEEE 10th Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Nov. 2011, pp. 847_852.

[11] X. Wu, X. Ying, K. Liu, and L. Chen, ``A survey of privacy-preservation of graphs and social networks,'' in *Managing and Mining Graph Data*. New York, NY, USA: Springer-Verlag, 2010, pp. 421_453

[12] M. Girvan and M. E. Newman, "Community structure in social and biological networks." *Proc Natl Acad Sci U S A*, vol. 99, no. 12, pp. 7821–7826, June 2002.

[13] Y. Wang, L. Xie, B. Zheng, and K. C. K. Lee, ``High utility k-anonymization for social network publishing,'' *Knowl. Inf. Syst.*, vol. 36, no. 1, pp. 1_29, 2013.

[14] L. Liu, J. Wang, J. Liu, and J. Zhang, "Privacy preservation in social networks with sensitive edge weights," in *Proc. 2009 SIAM Int. Conf. Data Mining (SDM'09)*, Sparks, NV, USA, Apr. 2009, pp. 954–965.

[15] A. Asuncion and D. J. Newman, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2010.

[16] J. Cheng, A. W. Fu, and J. Liu, "K-isomorphism: Privacy preserving network publication against structural attacks," in *Proc. 2010 Int. Conf. Management of Data (SIGMOD '10)*, New York, NY, USA, 2010, pp. 459–470, ACM.

[17] Shu, X, Yao, D & Bertino, E 2015, 'Privacy-preserving Detection of Sensitive Data Exposure', *IEEE Transactions on Information Forensics Security*, vol. 10, no. 5, pp. 1092-1103.

[18] Risk Based Security. (Feb. 2014). *Data Breach Quick-View: An Executive's Guide to 2013 Data Breach Trends*. [Online].Available:https://www.riskbasedsecurity.com/reports/2013- DataBreachQuickView.pdf, accessed Oct. 2014.

[19] M. O. Rabin, "Fingerprinting by random polynomials," Dept. Math., Hebrew Univ. Jerusalem, Jerusalem, Israel, Tech. Rep. TR-15-81, 1981.

[20] A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher, “Min-wise independent permutations,” *J. Comput. Syst. Sci.*, vol. 60, no. 3, pp. 630–659, 2000.

[21] M. Yuan and L. Chen, “Protecting Sensitive Labels in Social Network Data Anonymization ”,IEEE transaction on knowledge and data engineering, Vol. 25, No. 3,March 2013.

[22] J. Han, J. Pei, Y. Yin and R. Mao, *Mining frequent patterns without candidate generation: A frequentpattern tree approach*. Data mining and knowledge discovery, 2004. **8**(1): pp. 53-87.

Mitzenmacher, “Min-wise independent permutations,” *J. Comput. Syst. Sci.*, vol. 60, no. 3, pp. 630–659, 2000.

[21] M. Yuan and L. Chen, “Protecting Sensitive Labels in Social Network Data Anonymization ”,IEEE transaction on knowledge and data engineering, Vol. 25, No. 3,March 2013.

[22] J. Han, J. Pei, Y. Yin and R. Mao, *Mining frequent patterns without candidate generation: A frequentpattern tree approach*. Data mining and knowledge discovery, 2004. **8**(1): pp. 53-87.

Temitope B. Salau attained her B.Sc. degree in Computer Science from Niger Delta University, Nigeria in 2011. Recently obtained her M.Tech degree in Software Engineering from the School of Engineering in Sharda University, India in 2016. She has published a review paper on Privacy Preserving data mining which depicts different methods in which privacy can be preserved on data. She is currently making applications for PhD study.

Manoj K. Gupta received the B.Eng. degree in Computer Science and Engineering from CCS University, Meerut, India in 2001, the Master of Technology in Computer Science and Engineering from UPTU Lucknow, India in 2016 and Ph.D. degrees from the IIT Roorkee, Roorkee, India in 2014. He has about 4years research experience and 12years teaching experience. Currently, he is an Associate Professor in the Department of Computer Science and Engineering, Krishna Institute of Engineering & Technology, Ghaziabad, Uttar Pradesh, India. His research areas of interest include Bioinformatics, Algorithms, Computational biology and Data mining.

Abdullahi A. Haruna acquired his M.Tech degree in Software Engineering from Sharda University in 2015 and his B.Sc. degree in Computer Science from Kano University of Science and Technology in 2008. He acquired a diploma in Public administration from Federal College of Education, Kano in 2003. He also studied at the Emphatic Neoclassic College, Kano in 2006. Currently he is a part time lecturer at the Federal College of Education (FCE Kano) and also does Voluntarily Lecturing at Aminu College of Islamic and Legal Studies (Legal Kano). He has published three papers in Data Cleaning, Knowledge Discovery Approach to Repair a Textual Attributes in Data Mining, Association Rule Mining respectively.