International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064

# Advanced Unsupervised Anonymization Technique in Social Networks for Privacy Preservation

**Anjali R Kulkarni[1], Yogish H K[2]**

[1, 2]East West Institute of Technology, Bangalore, India

**Abstract:** ***We study the problem of anonymizing social network where the network data is split between several data holders or players. In such setting, each player controls some of the nodes and he knows only the edges that are adjacent to the nodes under his control. The goal is to provide anonymized view of entire network with respect to two scenarios. Scenario 1, where All players know the identities of all nodes but each player needs to protect the information from other players is the existence or non -existence of edges adjacent to his nodes. Scenario 2, each player needs to protect the identities of all nodes under his control along with the existence or non-existence of edges adjacent to his nodes. We start the study with sequential clustering algorithm applied to centralize and scenario 1 of distributed setting. Then we extend the algorithm to scenario 2 which is the most complicated part according to previous studies. Finally we conclude by outlining the future research proposals in that direction.***

**Keywords:** clustering, Social networks, anonymization, adversaries, coalition, descriptive data

## 1. Introduction

A social network is a social structure containing a set of individuals or organizations or even entire societies called social actors and a set of ties between these social factors that may be interactions or relationships. The term social network is used to describe a social structure determined by such ties. The tie through which any given social unit connects to other represents the convergence of the various social contacts of that unit. For example, a social network provides information on individuals in some population and the links between them, which may describe relations of friendship, collaboration, correspondence and so forth.

An information network, as another example, may describe scientific publications and their citation links. Networks are modeled by a graph, where the nodes of the graph correspond to the entities, while edges represent relations between them. Real social networks may be more complex or contain additional information. For example, in networks if the interaction is asymmetric the graph would be directed(e.g., A financial transaction network),if the interaction involves more than two parties then the network would be modeled as a hyper-graph(e.g., a social network that describes co-membership in social clubs); in case where there are several types of interaction, the edges would be labeled; or the nodes in the graphs are accompanied by attributes that provide demographic information such as age, gender, location or occupation which could shed light on the structure of the network.

The social network approach to understand social interaction is that social phenomena should be investigated through the properties of relations between and within units, instead of the properties of these units themselves. Thus, one common criticism of social network theory is that individual agency is often ignored although this may not be the case in practice. Precisely because many different types of relations, singular or in combination, form these network configurations, network analytics are useful to a broad range of research enterprises. The perspective of social network provides a set of methodologies for analyzing the structure of whole social entities as well as a variety of theories explaining the patterns observed in these structures. The study of these structures uses social network analysis to identify local and global patterns, locate influential entities, and examine network dynamics.

Analysis of social networks is an interdisciplinary field which emerged from social psychology, sociology, statistics, and graph theory. Social network analysis is now one of the major paradigms in contemporary sociology, and is also employed in a number of other social and formal sciences. Together with other complex networks, it forms part of the field of network science. Social network analysis deals with uncovering patterns in the connections between entities. It has been widely applied to organizational networks to classify the popularity or influence of individuals and to detect collusion and fraud. Social network analysis can also be applied to study disease transmission in communities, the functioning of computer networks, and emergent behavior of physical and biological systems.

Because of the technological advances it is easier to collect the electronic records that describe social networks. However, there will be two choices for agencies and researchers who collect data. Either they can publish data for others to analyze, even though it will create privacy threats, or because of privacy concerns they can withhold data, even though it leads to difficulty in the analysis of data. In on-line settings digital traces of human social interactions can be found in a wide variety, and this has made them rich sources of data for large-scale studies of social networks. While a number of these on-line data sources are based on blogging and social networking sites, where users have explicitly chosen to publish their links to others. Many of the most promising opportunities for the study of social networks are emerging from data on domains where users have strong expectations of privacy, these include e-mail and messaging networks, as well as the link structure of closed on-line communities. As a useful working example, consider a “communication graph,” in which nodes are e-mail addresses, and there is a directed edge (u, v) if u has sent at

least a certain number of e-mail messages or instant messages to v, or if v is included in u's address book.

There is a tremendous growth in the amount of personal data that can be collected and analyzed. For this to happen many of data mining tools are necessary with the aim to infer the trends to predict the future. However such a data should be protected against privacy a intrusion that restricts the direct access to personal information. But access to large amounts of data is essential in order to draw accurate inferences. For example hospitals may decide to collaborate in order to find out the effect of few diseases in the early stages. This requires access to patient's medical records, violating doctor-patient privilege. The remedy is to provide the data in a manner that enables one to draw inferences without violating privacy of individuals.

Data in social networks cannot be released to web as it is, since it might contain sensitive information. In order to address the need to respect the privacy of the individuals whose sensitive information is included in the data it is needed to anonymize the data prior to its publication. Data anonymization typically trades off with utility. Hence, it is required to find strategies in such that released anonymized data holds enough utility, as well as preserves privacy to some accepted degree.

## 2. Survey on Anonymization Techniques in Social Networks

The naive anonymization techniques used the method of removing identifying attributes like names or social security numbers from the data. The first attempt [1] in this regard is well-known problem of *k*-anonymization in the context of tables. By considering $E = \emptyset$ it totally suppresses the structural information, and the social network reduces to a collection of tabular records.

It considers the problem of releasing information from relational database that contains personal records ensuring individuals privacy as well as maintenance of data integrity. A release is considered k-Anonymous if the information for each person contained in the release cannot be distinguished from atleast k-1 other persons whose information also appears in the release. The main goal is to minimally suppress the cells in order to ensure that the released network is k-anonymous. This problem is NP-hard for ternary attribute values. It strengthens the NP-hardness that requires the domain name values to be larger than the number of tuples in the table. The algorithm used considers the 0 (k) approximations with arbitrary alphabet size which is based on graph representation. It provides an improvement over previously considered best approximation guarantee of 0(klogk).For binary alphabets the approximation factor achieved is 1.5 for k=2,and factor of 2 for k=3.But it is not possible to achieve an approximation better than k/4 using graph representation.

In designing studies of such systems, one needs to set up the data to protect the privacy of individual users while preserving the network properties. This is typically done through simple procedure in which each individual's name is replaced by a random user ID .For example e-mail address, phone number, or actual name, but the connections between the anonymized people are revealed, like who called to whom through call, who corresponded with whom, or who messaged to whom. The motivation behind anonymizing is, there may be considerable value in allowing researchers to study the structure of social network, while it is labeled with actual names and is sensitive and cannot be released. Researchers are not interested in "who" which is represented by each node, but in the properties of the graph, such as its connectivity, distances between nodes, frequencies subgraphs, or the extent to which it can be clustered. Anonymization here refers to exactly preserving the pure structure of the graph while suppressing the "who" information.

With respect to publically available databases, release of database after performing k-anonymity prevents definitive record linkages. It means in a k-anonymized graph, for each record in publically available database at least k-records could correspond to it, so that it hides each individual in a crowd of k records. Here the parameter k is chosen based on the privacy required in the particular application. One property that must be achieved in k-anonymity model is that, for each record in public database, all the corresponding records in k-anonymized graph must have same value for sensitive attribute. In order to achieve this property a constraint is added that specifies that sensitive attribute should take r distinct values for each cluster in the k-anonymized graph. It forms a direction for future research. Another direction may b extending k-anonymity to deal with changes in database. For example a hospital may be interested to release anonymized version of its patient's database on periodic bases. But it may lead to record linkages for some of the records as the several versions of anonymized database have been released which resulted in leakage of information. So in order to handle insert, delete and update operations to database the k-anonymity model can be extended.

The difficulty with this previous technique is that anonymous social network data almost never exists in the absence of outside context, so that an attacker can potentially combine this knowledge with the observed structure to begin compromising privacy, de-anonymizing nodes and even learning the edge relations between explicitly named (de-anonymized) individuals in the system. Moreover, such an adversary may in fact be a user (or set of users) of the system that is being anonymized.

In [2] families of attacks have been described which say that it is possible for an adversary to learn whether edges exist or not between a pair of target nodes even from a single anonymized copy of a social network.

The ways in which an adversary might take advantage of context is distinguished into two categories. These attacks are analogous to passive attacks and active attacks in cryptanalysis i.e. attacks in which an adversary simply observes data as it is presented, and those attacks in which the adversary tries to access the data to make it easier to decipher. In active attacks an adversary chooses an arbitrary set of users to violate their privacy, creates a number of new

user accounts with edges to these targeted users, and creates a pattern of links among the new accounts with the goal of making it stand out in the anonymized graph structure. Hence the adversary finds these new accounts together with the targeted users in the anonymized network that is released. That means in an n node network, if the attacker creates O(sqrt(log n)) nodes ,then it can begin compromising the privacy of arbitrary targeted nodes, with high probability for any network. Experiments show that, in a 4.4-million-node social network, the creation of 7 nodes by an adversary can compromise the privacy of about 2400 edge relations on an average. And the experiments also suggest that it is very difficult to determine whether a social network is compromised by such an active attack. In passive attacks, users do not create the new nodes or edges. Instead they try to find themselves in the released network, in order to know the existence of edges among users to whom they are linked. In the same social network containing 4.4-million-nodes, for the majority of users, it is possible to exchange structural information, and subsequently uniquely identify the subgraph on this coalition in the ambient network. With this, the coalition can then compromise the privacy of edges among pairs of neighboring nodes.

The active attack is structured by, creating k new user accounts (for some small parameter k), before the anonymized graph is produced and subgraph H is created by linking them together. Then using these accounts links are created (e.g.by sending messages) to nodes in {w1, . . . ,wb}, and as well as other nodes. So now, this subgraph H will be present when the anonymized copy of G is released, , as will the edges connecting H to w1, . . . ,wb. The attacker finds the copy of H that is present in G, and from this it locates w1, . . . ,wb. Thus the true location of these targeted users in G is identified, and the attacker can then determine all the edges among them, by compromising privacy.

Whereas, The passive attack assumes that most nodes in real social network data already belong to a small uniquely identifiable subgraph. Hence, if a user u is able to collude with a coalition of k − 1 friends after the release of the network, he or she will be able to identify additional nodes that are connected to this coalition, and hence learn the edge relations among them.

It may be passive or active attack; they do not have access to highly resolved data like time-stamps or other numerical attributes. They can only know about who links to whom, and not other node attributes, and hence this makes their task more challenging. Constructing the subgraph H, which involves hiding secret messages for later recovery using just the social structure of G, can be seen as a kind of structural steganography. Hence this approach can be seen as a step toward understanding how techniques of privacy-preserving data mining can inform how we think about the protection of even the most skeletal social network.

This work is not applicable to all settings in which social network data is used. Results show that one cannot rely on anonymization to ensure individual privacy in social network data, in the presence of parties that are trying to compromise this privacy. One way to achieve this is to try inventing methods of thwarting the particular attacks that are described, true safeguarding of privacy requires mathematical rigor, beginning with a clear description of what is meant by compromising privacy, to what information does it have access ,and what are the computational and behavioral capabilities of the adversary. In literature the problem of ensuring privacy is in settings such as work which rekindled interest in a field quiescent since the 1980s, and increasingly incorporating approaches from modern cryptography for describing information leakage. The notion of differential privacy gives very strong guarantees that are independent of the auxiliary information and computational powers of the adversary. This notion, instead of concentrating on how the database behaves with versus without the data of an individual, it compares what can be learned about an individual with versus without the database. The design of non-interactive mechanisms for ensuring reasonable notions of privacy in social network data is an open question, and potential results are constrained by these existing impossibility results. Hence, when computational safeguards are sought to protect social network data, the only techniques of which we are aware at the present time for simultaneously ensuring individual privacy and permitting accurate analysis, when the questions are not known in advance, are interactive.

Privacy cannot be guaranteed by simply removing the identities of the nodes before publishing the graph or social network data. The structure of the graph i.e. degree of the nodes, can reveal the identities of individuals. In order to address this issue, one needs to apply a more efficient procedure of anonymization on the network before it is released. There are 3 categories of privacy preservation in networks. The first category methods provide *k*-anonymity via edge additions or deletions, which is a deterministic procedure. In those methods assumptions are made that the adversary has background knowledge about some of the property of its target node, and then those methods modify the graph so that graph becomes *k*-anonymous with respect to that assumed property.

In order to prevent the disclosure of individuals identity, in [3] there is a framework called graph-anonymization .Given a graph *G* and an integer *k*, it then modifies *G* via a set of edge-addition or deletion operations to construct a new *k*-degree anonymous graph be *G*, so that in G every node *v* has the same degree with at least *k* -1 other nodes. It is possible to transform *G* to the complete graph, in which all nodes will be identical. Such an anonymization will preserve the privacy of individual nodes, but it makes the anonymized graph useless for other studies. Hence additional requirement such that the minimum number of edge modifications is made. In this way, both utility of the original graph is preserved, and at the same time degree-anonymity constraint is satisfied.

An assumption is made here that the graph is simple, *i.e.*, the graph is undirected, unweighted, containing no self-loops or multiple edges. Main focus is on the problem of edge additions, however the case of edge deletions is symmetric and thus can be handled analogously, for which it is enough to consider the complement of the input graph. The method to extend the proposed framework to allow simultaneous edge addition and deletion operations when modifying the input graph is also described.

A vector of integers V is said to be k-anonymous, if every distinct value in v appears at least k times. For example, vector v = [5, 5, 3, 3, 2, 2, 2] is said to be 2-anonymous. If the degree sequence of G, G', is k-anonymous then that graph G (V, E) is said to be k-degree anonymous,. This definition states that for every node v ЄV there exist at least k - 1 other nodes that have the same degree as v. With a priori knowledge of the degree of few nodes, this property prevents the re-identification of individuals by attacker.

It is proved that, for every k2 ≤k1, if a graph G(V,E) is k1-degree anonymous, then it is also k2-degree anonymous. Now the Graph Anonymization problem is defined with this. The input to the problem is a simple graph G(V,E) and an integer k. And a set of graph-modification operations are applied on G in order to construct a k-degree anonymous graph G'(V',E') that is structurally similar to G. We require that The output graph is seemed to be on same set of nodes as the original graph, that is, V' = V . Graph-modification operations are restricted to edge additions, that is, by adding a minimal set of edges graph G' is constructed from G. The cost of anonymizing G by constructing G' the graph-anonymization cost $G_a$ and we compute it as $G_a$( G',G) = |E'|-|E|. Graph Anonymization is defined as, given a graph G(V,E) and an integer k, and a k-degree anonymous graph G'(V,E') with 'E ˆ E = E such that Ga( G',G) is minimized.

The main problem is divided into two subproblems and an efficient algorithm for solving those subproblems is proposed. These algorithms can be applied to a set of and real-world graph data and the utility of the degree-anonymous graphs and its efficiency can be demonstrated. And also simultaneous edge additions and deletions can be performed by extending these algorithms.

Dealing with graphs when compared to existing data anonymization and perturbation techniques for tabular data, is a much more challenging task. In tabular data, each tuple can be considered as an independent sample from some distribution. But in a graph, all the nodes and edges are correlated; a single change of an edge or a node can alter the whole network. Moreover, with graphs it is difficult to model the capability of an attacker. In order to derive private information, the topological structure of the graph can be potentially used. Finally, it is considerably difficult to measure the utility of a graph. There are no any effective metrics to quantify the information loss incurred by the changes of its nodes and edges in the graph.

It is just an attempt to address some of these issues using simple and intuitive notions. Lots of additional work is needed in order to develop theoretically and practically efficient privacy models for graphs. The second category methods of privacy preservation add noise to the data, to prevent identification of their target in the network by attackers, or inferring the existence of links between nodes in the form of random additions, deletions or switching of edges. In [4] a framework for assessing the privacy risk of sharing anonymized network data is presented. A model of adversary knowledge, with several variants is considered and connections are made to known graph theoretical results. It shows that simple anonymization techniques are not adequate, resulting in privacy breaches for even modestly informed adversaries. Based on perturbing the network it proposes a novel anonymization technique and empirically demonstrates that it leads to reduction in privacy threat. It also analyzes the effect of anonymizing the network on utility of the data for social network analysis.

Here the social network is modeled as an undirected, unlabeled graph. The main objective of the data trustee is to publish the data in such a way that it permits useful analysis as well as protecting the privacy of entities represented. The first step in preparing the social network data for release is to remove identifying attributes such as name or social security number. Identity of nodes in the graph of relationships can be preserved by, giving names to synthetic identifiers. This procedure is named as the naive anonymization of a social network. Naive anonymization is a common practice, for example, in network trace data the identifying attribute is the IP address. Network traces are released after encrypting the IP address. Social network analysis can be performed in the absence of names and unique identifiers, thus Naive anonymization achieves utility goals of the data trustee. Here focus is on an attacker whose aim is to re-identify a known individual in the naively anonymized graph. Synthetic identifiers reveal nothing about node in the graph. But adversary may collect information from external sources about an individual's relationships, and may be able to re-identify individuals in the graph.

Thus, in the graph of relationships an entity's position acts as a quasi-identifier. The structural similarity of nodes in the graph, and the kind of background information an adversary can obtain decides, the extent to which an individual can be distinguished using graphical position.

It formalizes the re-identification threat and different kinds of adversary external information. Study includes a spectrum of outside information and shows the capacity to re-identify individuals in a graph. The threat of re-identification is related to results in random graph theory. These theoretical results are contrasted here with observations of re-identification attacks on real-world social networks. Protecting against the threat of re-identification presents challenges for graph structured data. In tabular data, identification of attributes can be generalized, randomized or suppressed easily, and their effects are largely restricted to the individual affected. It is not easy to generalize or perturb the structure around a node in a graph, and the impact of doing so can spread across the graph. We propose a novel alternative to naive anonymization based on random perturbation. Our perturbation techniques leave nodes unmodified but perform a sequence of random edge deletions and edge inserts. We show that this technique can significantly reduce the effectiveness of re-identification attacks by an adversary with acceptable distortion of the graph. We evaluate all our techniques on real datasets drawn from the domains mentioned previously: an organization social network derived from the Enron dataset, a network trace graphs from a major university, and a scientific collaboration network.

We have focused here on what we believe to be one of the most basic and distinctive challenges understanding the extent to which graph structure acts as an identifier and the

cost in accuracy required to obscure this identifier using perturbation. In this section we briefly describe alternatives to assumptions we have made and promising directions for future study.

We have also assumed the adversary targets one node at a time. That is, re-identification is focused on node x, and is considered independently of attempts to reidentify x0. Targeting sets of nodes simultaneously can have some subtle consequences. For example, if cand(x) = {y, y0} and cand(x0) = {y0} then x is uniquely identified since x0 can only correspond to y0. These overlapping but non-identical candidate sets are impossible for queries Hi (see Section 2). But they are possible for other knowledge queries that do not provide complete information. The general observation is that for some inference processes by the adversary, cand(x, x0) (the feasible assignments to the pair of targets x, x0) is not equal to cand(x)×cand(x0) (the cross product of the candidate sets for the individuals). Against an adversary seeking to re-identify a group of individuals, this aspect of the reasoning process must be taken into account.

The third category methods of anonymization do not alter the graph data as in the methods of the two previous categories. Instead, the nodes are clustered together into super-nodes of size at least *k*, where *k* is the required anonymity parameter, and then publishing the graph data in that coarse resolution. The problem of *k*-anonymization of social networks by clustering was considered by the studies done by Zheleva and Getoor [5].Here the main idea was, arriving at a clustering of the nodes, by applying any standard *k*-anonymization algorithm on the quasi-identifier records that describe the nodes. And here five ways are suggested to hide the structural information.

The focus in this study is on preserving the privacy of sensitive relationships in graph data. We refer to The problem of inferring sensitive relationships from anonymized graph data is referred here as link re-identification. For this, five different privacy preservation strategies are proposed, that vary based on the amount of removed data (and hence their utility) and also the amount of privacy that is preserved. An assumption is made that is, adversary has an accurate predictive model for links, and the success of different link re-identification methods under varying structural characteristics of the data has been shown experimentally.

The graph data describing entities and relationships between entities is considered. Relationships are assumed to be binary relationships. As usually, in a graph, each entity is denoted by node, and relationship is denoted by edge. In general, there can be different types of nodes and different types of edges. In this study, the focus is on the case where there is a single node type and multiple edge types. One of the relationship types is distinguished as the sensitive relation-ship. And his sensitive relationship needed to be hidden from the adversary. In addition the nodes and edges can have associated attributes, as well as the graph has structural properties. Structural properties of a node are degree of the node and structure of neighborhood.

The process of anonymization in this study is described as, taking the unanonymized graph data, making some modifications to it, and construction of a new released graph which will be available to the adversary. The modifications include changes to both the graph nodes and its edges. Several graph anonymization strategies have been discussed in this study and, later for each approach, the tradeoffs between utility and the privacy preservation of the anonymized data is discussed.

In node anonymization techniqu, it is assumed that, anonymization of nodes is done with one of the techniques used for single table data. For example, using t-closeness the nodes could be k-anonymized. This anonymization results in clustering of the nodes into m equivalence classes as ($C_1$, . . . ,$C_m$) such that ,in its quasi-identifier attributes ,each node is indistinguishable from some minimum number of other nodes. The anonymization of nodes gives raise to equivalent classes of nodes. Inside each of this equivalence class, there can be nodes with different identifying edges and structural properties; hence these resultant equivalent classes are based only on node attributes.

In order to describe relational part of the graph, five possible anonymization approaches have been defined. They range from one which removes very less amount of information to a very restricted one that removes the greatest amount of relational data. The first (trivial) edge anonymization approach involves, leaving all other observational edges intact only removing the sensitive edges. Hence this method is called intact edge anonymization. This anonymization technique should have a high utility, as the relational observations remain in the graph. But this approach is likely to have low privacy preservation. Another anonymization approach involves removing some portion of the relational observations. Hence this method is referred as partial edge removal anonymization. A particular type of observation can be removed, which contributes to the overall likelihood of a sensitive relationship, or a certain percentage of observations can be removed that meets some pre-specified criteria. This partial edge removal approach should decrease the utility of the data and increase the privacy preservation as compared to the previous method. Removal of observations should decrease the number of node pairs that are likely having sensitive relationships but it will not remove them completely. Private information for those pairs of nodes, may be disclosed. The simplest approach is to leave the sets of edges intact, and for each edge type, maintaining the counts of number of edges between the clusters. This technique is referred to as cluster-edge anonymization. Next, a very stricter method as compared to previous methods is considered for sanitizing observed edges that is cluster-edge anonymization with constraints technique.

It creates edges between equivalence classes as in cluster edge-anonymization, but it needs the equivalence class nodes to have the same constraints as any two nodes in the original data. This results in removal of some of the count information which is revealed in the previous anonymization technique. In literature there are two types of privacy attacks of data those are identity disclosure and attribute disclosure. In graph data, there is a third type of attack called as link re-identification. Link re-identification infers that two entities participate in a particular type of sensitive relationship or communication. An adversary can make Sensitive

conclusions about the data that are more general statements, and can involve node, edge and structural information. These conclusions can be the results of aggregate queries.

Because understanding and appreciating the effectiveness of techniques is such an important and timely topic, this work will motivate further research in the literature. In literature, to address both descriptive and structural data, Campan and Truta [6] were the first to apply an anonymization algorithm. The algorithm proposed by them, was analogous to SaNGreeA (Social Network Greedy Anonymization), greedy clustering algorithm, generating one cluster at a time, by initially selecting a seed node and then keep on adding next node to it,such that it minimizes some information loss measure, until it produces a cluster of size *k*. As this algorithm builds the clustering gradually, the actual information loss measure $I(\cdot)$ cannot be used. The structural information loss, $IS(\cdot)$ can be evaluated once when the clustering is defined. For this reason they replaced $IS(\cdot)$ with a distance metric between nodes, and it was experimentally an effective substitute. The sequential clustering algorithm that was presented in this study does not suffer from those problems. Because in each stage of its execution it has a full clustering and hence it may always make decisions according to the real measure of information loss. The main focuses in this paper are a greedy algorithm for anonymization and a measure to quantify the information loss in the anonymization process due to edge generalization.

In this study, a new anonymization approach is proposed for social network data that consists of nodes and relationships. A node is an individual entity and is described by identifier attribute (such as Name and SSN*)*, quasi-identifier attribute (such as ZipCode and Sex*)*, and sensitive attributes (such as Diagnosis and Income) . In between two nodes a relationship exists and it is unlabeled, that means all relationships have the same meaning. Masking is done in order to protect the social network data, with respect to the *k*-anonymity model and it says that every node will be indistinguishable in terms of both node's attributes and structural information with at least (*k*-1) other nodes. This anonymization method tries to disturb the social network data as little as possible, both the attribute data associated to the nodes, and the structural information. The method used here for attribute data anonymization is generalization. The method used for structural anonymization,is called edge generalization. It does not add or remove edges from the social network dataset. To quantify the amount of information loss caused by edge generalization through cluster collapsing an information loss measure is defined. The cluster formation process defined here gives more importance to the nodes' attribute data and equally to the nodes' neighborhoods. This process is user-balanced as it preserves more structural information of the network, as measured by the structural information loss, and the nodes' attribute values, which are quantified by the generalization information loss measure.

In this study, the social network privacy model is defined as a simple undirected graph G = (N,E),where N is the set of nodes and $E \subseteq N \times N$ is the set of edges. Each individual entity is represented by node. Each relationship between two entities is defined by edges. The set of nodes is described by a set of attributes that are classified into the three categories. Identifier attributes such as Name and SSN that can be used to identify an entity. Quasi-identifier attributes such as Zip code and Sex that may be known by an intruder. Confidential or sensitive attributes such as Diagnosis and Income that are assumed to be unknown to an intruder.

In this model only binary relationships are assumed. All relationships are considered to be of same type, so they are represented by unlabeled undirected edges. This type of relationship is of same nature as all the other quasi-identifier attributes. This type of relationship is referred as quasi-identifier relationship. It means that the graph structure may be known to an intruder and he may match it to known external structural information, therefore resulting in to privacy attacks that might lead to identity and/or attribute disclosure.

In this study a technique called generalization of the quasi-identifier attributes is defined, which is widely used for microdata k-anonymization. It involves replacing the actual value of an attribute with a more general n less specific value that is equivalent to the original. This technique is reused for the generalization of nodes attributes' values.

To quantify the structural information a measure is defined, when anonymizing a graph through collapsing clusters into nodes, together with their neighborhoods, it is lost. And the Information loss quantifies the probability of error generated while trying to reconstruct the structure of the initial social network from its masked version. There are two components for the structural information loss i.e. the intra-cluster structural information loss and the inter-cluster structural information loss components.

The algorithm described in this study is called the SaNGreeA (Social Network Greedy Anonymization) algorithm, to generate a *k*-anonymous masked social network; it performs a greedy clustering processing. Here the given social network is modeled as a graph $G = (N, E)$. Nodes are described by quasi-identifier and sensitive attributes and edges are undirected and unlabeled. First, the algorithm partitions N nodes into clusters. Next, within each cluster, all the nodes are made uniform with respect to the quasi-identifier attributes and relationship. This is achieved by using generalization, both for the quasi-identifier attributes and the quasi-identifier relationship. In order to meet the requirements of the *k*-anonymity model, each cluster has to contain at least *k* tuples. Hence, a first criterion in clustering process is to ensure that each cluster has enough elements. In order to minimize the information lost between the initial social network data and its masked version, caused by the subsequent cluster-level quasi-identifier attributes and relationship generalization a second criterion is used. The clustering algorithm uses two information loss measures in order to obtain good quality, as well as to permit the user to control the type and the quantity of information loss. To quantify how much *descriptive* data detail is lost through quasi-identifier attributes generalization, a metric is used called as generalization information loss measure. The second measure quantifies how much *structural* detail is lost through the quasi-identifier relationship generalization and it is called structural information loss.

Because SaNGreeA is a greedy algorithm which selects solution from search space based on local optima of two measures, it finds a good solution to anonymization,but not the best solution among the existing ones.The time complexity of SaNGreeA is $O(n^2)$.The method to find efficient solution is not known. The k-anonymization for microdata is found to be NP-hard, and the optimization problem defined here is same with the only difference is minimizing two information loss measures before release of data.

This study gives way to researches in several directions. such as extending the anonymity model to achieve protection for disclosure in social networks. Or studying the change in the utility of anonymized social network for various application fields.

We study the problem of privacy-preservation in social networks in [7], that considers the distributed setting in which the network data is split between several players. The goal is to arrive at an anonymized view of the unified network without disclosing information about links between nodes to any of the data holders that are controlled by other data holders. Here the study starts with the centralized setting and based on sequential clustering two variants of an anonymization algorithm are offered. The performance of these algorithms is significantly higher than the SaNGreeA algorithm due to Campan and Truta which was the leading algorithm in literature for achieving anonymity in networks by means of clustering. Then the secure distributed version of these algorithms is devised. This study is first for privacy preservation in distributed social networks.

In this study the social networks where the nodes are accompanied by descriptive data are considered, and two novel anonymization methods of the third category anonymization(namely, by clustering the nodes)are proposed. These algorithms issue anonymized views of the graph with significantly smaller information losses than anonymizations issued by the previous algorithms in literature. And also the distributed versions of algorithms are proposed and their privacy and communication complexity are analyzed.

Anonymization here is done by sequential clustering. For $k$-anonymizing tables, the sequential clustering algorithm was found to be a very efficient algorithm with respect to runtime as well as the utility of the output anonymization. Here an adaptation of it for anonymizing social networks is done. Algorithm for centralized setting starts with a random partitioning of the network nodes into clusters. The initial number of clusters in the random partition is set to $\lfloor N/k0 \rfloor$. And all of the initial clusters are of size $k0$ or $k0 + 1$, where $k0 = \alpha k$ is an integer and $\alpha$ is some parameter that needs to be determined. The algorithm then starts its main loop. In the main loop, the algorithm goes the $N$ nodes in a cyclic manner and for each cycle for each node it checks whether that node can be moved from its current cluster to another one with less information loss of the resultant anonymization. If such an improvement is possible over the information loss, the node is transferred to the cluster where it fits best currently. Some of the clusters may be large at this point, i.e. their size. is at least $k$, while others are small. We apply an agglomerative procedure if there exist small clusters. At least one of them is selected at random and an agglomerative procedure is applied on them and then which of the other clusters (of any size) is closest to it is determined, unifying them will cause the smallest increase in the information loss. After finding the closest cluster, the two clusters are unified. This procedure is repeated until all clusters are of size $k$. The parameters $\alpha$ and $\beta$ control the sizes of the clusters and, that means information loss of the final output. The goal is to find a setting of $\alpha$ and $\beta$ such that they yield lower information losses.

Modified structural information loss is described in this study. The SaNGreeA algorithm uses a measure of structural information loss that differs from the measure $IS(\cdot)$ .It is redefined here. In other words, $I'_S$ of a given cluster is the average distance between all pairs of nodes in that cluster, and $I'S$ of the whole clustering is the corresponding weighted average of structural information losses over all clusters. The significant difference between $I(\cdot)$ and $I'(\cdot)$ is that the former cannot be evaluated until the entire clustering is determined, while the latter one is defined as a sum of independent intra-cluster information loss measures, it can. As the SaNGreeA algorithm uses a distance function between a cluster and node that is geared towards minimizing the measure $I'(\cdot)$ .Hence it needs to make clustering decisions before all clusters are formed. The sequential clustering algorithm can use either $I(\cdot)$ or $I'(\cdot)$.

For distributed setting, network data is split among $M$ sites (or players), and there are two scenarios to consider in this setting:

1) Scenario A: Each player needs to protect the identities of the nodes that are in under his control from other players, and also he has to protect the existence or non-existence of edges adjacent to his nodes.
2) Scenario B: All players know the identities of all nodes in V; the information that each player has to protect is the existence or non-existence of edges adjacent to his nodes from other players.

In this study focus is on Scenario B. Scenario A is significantly harder and is left for future research. Distributed version of the sequential clustering is presented here, that uses the modified information loss measure, $I'(C)$. The distributed sequential clustering which is guided by the original information loss measure, $I(C)$ also works similar. In Scenario B, the descriptive information of all nodes can be made known to all players. Hence, the difference in the descriptive information loss, $ID(\cdot)$, can be computed openly not securely. It is the difference in the structural information loss, $I'S(\cdot)$.As it depends on the edge structure of the graph which is split between the various players and must not be disclosed, it must be computed in a secure manner. A secure multi-party protocol (SMP hereinafter) that performs such computations is defined.

For distributed setting, the three main stages of Algorithm for centralized setting is revisited and its implementation in secure manner is explained. First step is initial partitioning, in which each player generates a uniform and random labeling of his own nodes by labels from $\{1, \ldots, T := \lfloor N/k0 \rfloor\}$. The cluster Ct, $1 \leq t \leq T$, consists of all nodes in $V$ that has the label $t$. This allocation of nodes to each cluster is

made known to all players. Second step is Single node transitions ,during the main loop in algorithm, difference in the structural information loss if a given node *Vn* would move from its current cluster *Ct* to any of the other clusters, Cs, s ∈ [*T*] \ {t} is computed. Third stage is, the agglomerative stage, *here* change in I′*S* if clusters *Ct* and *Cs* would be unified.

Computing the sum of private integers has well known simple SMPs where the private vectors are added. The components of the vectors are rational numbers. The denominators of those rational numbers are common and known to all, but the numerators depend on private integers. Hence, it is the problem of computing sums of private vectors over the integers. It is possible to compute upfront an upper bound *p* on the size of those integers and their sum. Thus the problem may be further reduced to computing sums of private vectors.

It is proved here that, each sequence of clustering's that can be realized during an *M*-distributed implementation of Algorithm for centralized setting on given inputs, is a possible sequence also in a centralized implementation (*M* = 1), and vice-versa.

This study provides many directions for future researches. One direction this study suggests is to devise distributed algorithms also to Scenario A which is not addressed here. In that scenario, each of the players needs to protect the identity of the nodes under his control from the other players. Hence, it is more difficult than Scenario B .As it requires a secure computation of the descriptive information loss (while in Scenario B such a computation can be made openly in a public); and the players must hide from other players the allocation of their nodes to clusters. Another research direction that this study suggests is to devise distributed versions of the *k*-anonymity algorithm using different techniques.

## 3. Conclusions

We presented the study on what a social network is, how it is structured and modeled. And the need of social networks for researchers from various disciplines for study is depicted. Necessity of Data anonymization in such social networks prior to its publication in order to preserve the privacy. How the trade- off between data anonymization and utility poses challenge to researchers in this direction is explained. We studied various attempts in literature to achieve utility in one hand and to preserve the privacy to some accepted degree on other hand. For each of the study in this direction, the aims set, various methodologies used, and algorithms designed, mathematical formulations that have been applied, and the way they concluded their study with a way for future research directions are presented here. And to overcome the limitations posed by these studies, a new anonymization method is proposed based on sequential clustering.

## References

[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R.Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In ICDT, volume 3363 of LNCS, pages 246–258, 2005.
[2] L. Backstrom, C. Dwork, and J. M. Kleinberg.Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In WWW, pages 181–190, 2007.
[3] K. Liu and E. Terzi. Towards identity anonymization on graphs. In SIGMOD Conference, pages 93–106, 2008.
[4] M. Hay, G. Miklau, D. Jensen, P. Weis, and S.Srivastava. Anonymizing social networks. Uni. of Massachusetts Technical Report, 07(19), 2007.
[5] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationship in graph data. In *PinKDD*, pages 153–171, 2007.
[6] A. Campan and T. M. Truta. Data and structural k anonymity in social networks. In PinKDD, pages 33–54, 2008.
[7] Tamir Tassa and Dror J.Cohen.Anonymization of Centralized and Distributed Social Networks by Sequential Clustering.The Open University, Ra'anana, Israel.pages 1-14, 2013.

## Author Profile

**Anjali R Kulkarni** received the B.E degree in Computer Science and Engineering from Visvesvaraya Technological University in 2011. She is currently pursuing her M.Tech degree in Computer Science and Engineering under the same University.

**Mr Yogish H K** received the B.E and M.Tech degrees in Computer Science and Engineering from Visvesvaraya Technological University in 1997 and 2004, respectively. He worked as a Lecturer in Kalpataru Institute of Technology Tiptur during 1997-2002,and in Don Bosco Institute of Technology Bangalore during 2004-2005.Then he worked in Reva Institute of Technology and Management Bangalore as Lecturer during 2005-2007,and as Associate Professor during 2011-2012.Presently Working as Associate professor in Computer Science Dept East West Institute of Technology Bangalore.