XColor: Protecting General Proximity Privacy

***Abstract*— As a severe threat in anonymized data publication, proximity breach is gaining increasing attention. Such breach occurs when an attacker learns with high confidence that the sensitive value of a targeted victim belongs to a set of semantically proximate values, even though not sure about the exact value. Focusing on tackling proximity breach under specific data models (e.g., categorical or numeric sensitive data), existing research efforts fail to address the threat for a much richer set of models wherein the semantic proximity might be defined by arbitrary functions. This work advocates studying proximity privacy in a data-model-neutral manner, thereby providing solutions of general applicability. Specifically, a) we define the association between quasi-identifier and sensitive attributes (QI-SA association) under a highly abstract data model; b) we formalize proximity breach in the general framework of proximate QI-SA association, and remedy such breach with a unified privacy definition; c) for the proposed definition, we conduct an analytical study on its characteristics, and derive criteria to efficiently test its satisfiability; d) further, we devise a novel anonymization model,** XCOLOR**, to fulfill this definition, with guarantees on both operation efficiency and utility preservation. A comprehensive empirical evaluation is performed to validate the analytical models and the efficacy of the** XCOLOR **method.**

I. INTRODUCTION

Privacy preservation has become a paramount concern in numerous data dissemination applications that involve private personal information, e.g., medical data and census data. Typically, such *microdata* is stored in a relational table T: each record in T corresponds to an individual; the attributes of T are categorized as either *sensitive* or *non-sensitive*. In the setting of *central publication*, a publisher intends to release an *anonymized* version T^* of the microdata table T, such that no malicious user, called an *attacker*, can infer the sensitive information regarding any individual from T^*, whereas the statistical utility of T is still preserved in T^*.

Towards this end, a bulk of work has been done on anonymized data publication [3], [15], [16], [17], [18], [19], [21], [23], [25], [29], [28], [30]. One of the major aims is to address *association attack*: the attacker possesses the exact non-sensitive (*quasi-identifier* (QI)) values of the victim, and attempts to discover his/her sensitive (SA) value from the published table T^*. A popular methodology of thwarting such attacks is *generalization* [23]: after partitioning the microdata table T into a set of disjoint subsets of tuples, called *QI-group*, generalization transforms the QI-values in each group to a uniform format such that all tuples belonging to the same group are indistinguishable in terms of their QI-values.

Example 1. Consider publishing the medical data as shown in Table I: *age* and *zip-code* are QI-attributes, while *syndrome* is a composite SA-attribute, each component indicating the severity of a patient's suffering the corresponding symptom. The generalization of the microdata produces two QI-groups, as indicated by the group identifiers (GID). An attacker who knows *Alice*'s QI-values can no longer uniquely identify her SA-value: any tuple in the first group may belong to her; without further information, the attacker can only conclude that *Alice* associates with each specific *syndrome* value with identical probability 20%.

A. State of The Art

Essentially, the attack above is performed by leveraging the association between quasi-identifier and sensitive attributes (QI-SA association) appearing in the published data. Generalization weakens such association by reducing the representation granularity of QI-values. The protection is sufficient if the weakened association is no longer informative enough for the attacker to infer individuals' SA-values with high confidence. To gauge the quality of protection, a plethora of generalization principles have been proposed, which can be classified according to their targeted types of QI-SA association, namely, *exact* and *proximate* association.

Exact QI-SA association refers to the links between specific QI-values and SA-values in the published data. Exact association is particularly meaningful for publishing categorical sensitive data wherein different values tend to have no sense of proximity; it is desired to prevent the attacker from linking the victim to a specific SA-value with high confidence. The principles in this category include k-anonymity [23], l-diversity [18], and its variants [25], [29], all with the objective of avoiding uncovering exact SA-values.

Proximate QI-SA association, meanwhile, refers to the links between specific QI-values and a set of proximate SA-values. Clearly, this concept generalizes exact QI-SA association in that it takes account of the semantic proximity among SA-values. After studying the published data, even though with low certainty about the exact value, the attacker might have learned with high confidence that the SA-value of the victim belongs to a set of proximate values. A number of principles, including (k, e)-anonymity [30], variance control [15], and (ϵ, m)-anonymity [17], have been proposed to address such *proximity breach* under the model of one-dimensional numeric sensitive data (different values are strictly ordered).

B. Motivation

While focusing on tackling proximity breach under specific data models, be it categorical or numeric sensitive data, existing research efforts, however, fail to address the threat for a much richer set of models wherein the semantic proximity might be defined by arbitrarily complex or customized functions. An example is given as follows.

Example 2. Recall the running example of Table I. Assume that the semantic distance between two SA-values, represented

	age	zip-code	syndrome			GID
			allergy	asthma	myocarditis	
Alice 1	[18, 30]	[12k, 17k]	**0.8**	**0.0**	**0.0**	1
2	[18, 30]	[12k, 17k]	0.6	0.4	0.4	1
3	[18, 30]	[12k, 17k]	**0.7**	**0.1**	**0.1**	1
4	[18, 30]	[12k, 17k]	1.0	0.2	0.2	1
5	[18, 30]	[12k, 17k]	0.1	0.9	0.9	1
6	[32, 40]	[22k, 30k]	0.2	0.5	0.2	2
7	[32, 40]	[22k, 30k]	0.8	0.1	0.9	2
8	[32, 40]	[22k, 30k]	0.4	0.3	0.5	2
9	[32, 40]	[22k, 30k]	0.6	0.9	0.3	2
10	[32, 40]	[22k, 30k]	1.0	0.7	0.7	2

TABLE I: Anonymized data publication.

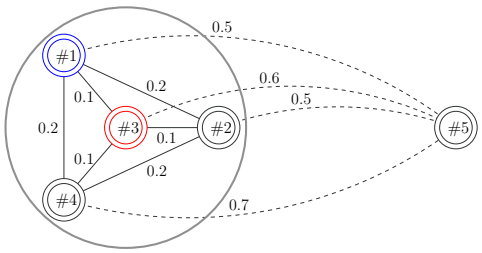

Fig. 1: Illustration of general proximity breach.

as two vectors $P = \langle p_i \rangle_{i=1}^n$ and $Q = \langle q_i \rangle_{i=1}^n$, is defined as $\Delta(P, Q) = \min_i |p_i - q_i|$. Measuring the pairwise distance of the *syndrome* values appearing in the first QI-group, one can notice that the first four tuples form a compact "neighborhood" structure, wherein the value of #3 is semantically proximate to that of #1, #2, and #4, as shown in Fig. 1.

From the attacker's perspective, every tuple in this group belongs to *Alice* with equal possibility; she can thus conclude that *Alice* associates with the neighborhood structure with probability 80%. Moreover, she might choose the value of the center node (#3) as an estimation, and arrives at a privacy intruding claim that "*Alice*'s *syndrome* value is fairly close to $(0.7, 0.1, 0.1)$".

Existing privacy principles and definitions, however, are incapable of capturing this general form of proximity breach because of their assumptions regarding the underlying data models. For example, in Table I, the SA-values in each QI-group are all distinct, thereby satisfying l-diversity [18] with the maximum possible $l = 5$; meanwhile, these multi-dimensional values can not be strictly ordered, thus rendering the techniques developed in [17] inapplicable.

Therefore, in this paper, we advocate studying proximity breach in a data-model-neutral manner, which we refer to as *general proximity breach*, with the objective of providing 1) a better understanding regarding proximity privacy and 2) anonymization solutions of general applicability. We argue that the proximity breaches addressed in the literatures, e.g., *homogeneity breach* [18], are essentially instantiations of this concept. In this paper, we aim at developing effective countermeasures to tackle such general breach.

It is worth contrasting our targeted setting with that of multiple sensitive attributes. We focus our discussion on the case of a single sensitive attribute, which might comprise multiple components (e.g., Table I), but is associated with a unified distance metric; while in the setting of multiple sensitive attributes, each attribute might be associated with a different distance metric, and possibly no single measure exists to capture the overall proximity. Following existing practices such as [18], our results can be readily extended to the case of multiple sensitive attributes.

C. Contributions

To the best of our knowledge, this paper presents the first systematic study on proximity privacy in a data-model-neutral manner, with findings of general applicability.

Concretely, we define QI-SA association under a highly abstract data model, with the only assumption of a semantic distance metric over the domain of the SA-attribute; then, we formalize general proximity breach in the framework of QI-SA association, and address such breach with a unified privacy definition, (ϵ, δ)-*dissimilarity*. It intuitively requires that in each QI-group, every SA-value should be "dissimilar" to a sufficient number of others.

We provide sound theoretical proof that (ϵ, δ)-dissimilarity, used in conjunction with k-anonymity [23] (called $(\epsilon, \delta)^k$-dissimilarity), offers effective protection against association attack, in terms of both exact and proximate QI-SA association. We further show that a number of existing privacy definitions are essentially instantiations of this general principle under the data models that they are designed for.

We conduct an analytical study on the characteristics of $(\epsilon, \delta)^k$-dissimilarity, derive criteria enabling to efficiently test its satisfiability for given microdata, and discuss the optimal setting of the parameters ϵ, δ, and k to achieve the best quality of privacy protection and utility preservation.

Most importantly, we propose a novel anonymization model, XCOLOR, to fulfill $(\epsilon, \delta)^k$-dissimilarity, with guarantees on both operation efficiency and utility preservation, by extending the techniques of *defect graph coloring*. Extensive experiments are conducted over real data to validate the practical performance of XCOLOR.

II. FORMALIZATION

In this section, we clarify the concept of general proximity breach, and prove the effectiveness of $(\epsilon, \delta)^k$-dissimilarity in remedying such breaches. Furthermore, we discuss the relevance of $(\epsilon, \delta)^k$-dissimilarity to existing privacy principles against proximity breach.

A. Models and Assumptions

Let T denote a microdata table intended to be published, which consists of d quasi-identifier (QI) attributes $\{A_i^{qi}\}_{i=1}^d$ and a sensitive (SA) attribute A^s. Particularly, 1) A^s can be of arbitrary data type, e.g., categorical, numeric, and customized defined type; 2) a semantic distance metric $\Delta(\cdot, \cdot)$ is defined over the domain of A^s, with $\Delta(x, y)$ denoting the distance between two SA-values x and y.

We begin with formalizing the concept of generalization:

Definition 1 (QI GROUP/PARTITION)**.** *The microdata table T is divided into m disjoint subsets of tuples $\mathcal{G}_T = \{G_i\}_{i=1}^m$,*

which satisfy (i) $\bigcup_{i=1}^{m} G_i = T$ and (ii) $G_i \cap G_j = \emptyset$ for $i \neq j$. Each G_i is called a QI-group, and $\mathcal{G}_T$ is referred to as a partition of T.

Definition 2 (GENERALIZATION)**.** *Given a partition $\mathcal{G}_T = \{G_i\}_{i=1}^{m}$, a generalization of T is a table T^* that is obtained by transforming the QI-values in each group G_i to a uniform format, i.e., all tuples in the same QI-group are indistinguishable with respect to their QI-values.*

For a numeric QI-attribute A^{qi}, the generalized value could be the minimum bounding interval of all A^{qi} values in the group; while for a categorical attribute A^{qi}, it could be the lowest common ancestor (LCA) of all A^{qi} values in the group on the domain generalization taxonomy of A^{qi}.

Further, we introduce another fundamental concept underlying the attack model, the neighborhood of a SA-value.

Definition 3 (ϵ-NEIGHBORHOOD)**.** *In a QI-group G with SA-values as a multi-set $\mathcal{SV}_G = \{v_i\}_{i=1}^{n}$, the ϵ-neighborhood of a value $v \in \mathcal{SV}_G$, $\Phi_G(v, \epsilon)$, is defined as the subset of $\mathcal{SV}_G$ with their distance to v within ϵ, formally*

$$\Phi_G(v, \epsilon) = \{v' |\ v' \in \mathcal{SV}_G \text{ and } \Delta(v, v') \leq \epsilon\}$$

Example 3. In the running example of Fig. 1, given ϵ = 0.1, the ϵ-neighborhood of v_3 consists of $\{v_1, v_2, v_3, v_4\}$.

We proceed to presenting the attack model. The attacker attempts to exploit the generalized table T^* to infer the SA-value $o.A^s$ of a targeted individual o. We assume that she possesses full identification information [19]:

Definition 4 (BACKGROUND KNOWLEDGE)**.** *The attacker possesses the information including (i) the exact QI-values of o, (ii) the QI-group G in T^* which o belongs to, and (iii) the semantic distance metric $\Delta(\cdot, \cdot)$ over A^s.*

By assuming the background knowledge (ii), we are dealing with the worst-case scenario that only one QI-group matches the QI-value of o in the generalized table T^*.

B. General Proximity Breach

After identifying the QI-group G containing o, the attacker proceeds to estimating $o.A^s$ following a probabilistic model:

Definition 5 (ATTACK MODEL)**.** *From the attacker's perspective, every tuple in G belongs to o with identical possibility; therefore, the probability that $o.A^s$ belongs to the ϵ-neighborhood of a SA-value v can be formulated as:*

$$\mathsf{prob}[o.A^s \in \Phi_G(v, \epsilon)] = |\Phi_G(v, \epsilon)|/|G| \tag{1}$$

where $|\Phi_G(v, \epsilon)|$ denotes the cardinality of the neighborhood.

It is clear now that exact QI-SA association is a special case of proximate QI-SA association under the setting of ϵ = 0; for any v, $\mathsf{prob}[o.A^s = v] \leq \mathsf{prob}[o.A^s \in \Phi_G(v, \epsilon)]$ for any $\epsilon > 0$.

Next, we formalize the concept of general proximity breach. Intuitively, if the ϵ-neighborhood $\Phi_G(v, \epsilon)$ encompasses a considerable proportion of the SA-values in G, the attacker can conclude that the victim o is associated with the SA-values appearing in $\Phi_G(v, \epsilon)$ with high probability, though she may not be sure about the exact value. Furthermore, by choosing v as the representative, she can arrive at fairly precise estimation about $o.A^s$, if ϵ is sufficiently small.

To measure the severeness of the privacy threats, and particularly, to capture the impact of proximate SA-values on enhancing the attacker's estimation, we introduce the metric of *proximity risk*.

Definition 6 (PROXIMITY RISK)**.** *Given the neighborhood radius ϵ, the risk of general proximity breach of a QI-group G, $\mathsf{risk}(G, \epsilon)$, is formulated as:*

$$\mathsf{risk}(G, \epsilon) = \max_{v \in \mathcal{SV}_G} \frac{|\Phi_G(v, \epsilon)| - 1}{|G| - 1} \tag{2}$$

Intuitively, $\mathsf{risk}(G, \epsilon)$ measures the relative size of the largest ϵ-neighborhood in G; by excluding v from the neighborhood, it highlights the effect of proximate SA-values on improving the attacker's belief: she priorly associates the victim with each SA-value[1] with identical probability $1/|G|$.

We note that $\mathsf{risk}(G, \epsilon)$ is a real number within the interval $[0, 1]$. In particular, G is free of proximity breach ($\mathsf{risk}(G, \epsilon)$ = 0) if all the SA-values are dissimilar, and reaches its maximum ($\mathsf{risk}(G, \epsilon)$ = 1) if a SA-value v is proximate to all other SA-values. Specially, we define that $\mathsf{risk}(G, \epsilon) = 1$ for the extreme case of $|G| = 1$.

Furthermore, we define the risk of general proximity breach, $\mathsf{risk}(\mathcal{G}_T, \epsilon)$, for a partition $\mathcal{G}_T$ of the microdata table T, as the maximum risk of all the QI-groups in $\mathcal{G}_T$, formally

$$\mathsf{risk}(\mathcal{G}_T, \epsilon) = \max_{G \in \mathcal{G}_T} \mathsf{risk}(G, \epsilon) \tag{3}$$

C. (Epsilon, Delta)^K-Dissimilarity

To remedy general proximity breach, we propose a novel privacy definition, (ϵ, δ)-*dissimilarity*.

Definition 7 ((ϵ, δ)-DISSIMILARITY)**.** *A partition $\mathcal{G}_T$ is said to satisfy (ϵ, δ)-dissimilarity if for each $G \in \mathcal{G}_T$, every SA-value v in G has less than $(1 - \delta) \cdot (|G| - 1)$ ϵ-neighbors.*

Here the parameter ϵ specifies the threshold of semantic proximity; while the parameter δ essentially controls the risk of potential proximity breach.

Next, we prove the effectiveness of this definition against association attack. Concretely, we show that a partition $\mathcal{G}_T$ is free of general proximity breach if and only if it satisfies (ϵ, δ)-dissimilarity. We have the following theorem:

Theorem 1. *Given the microdata table T and the neighborhood radius ϵ, for a partition $\mathcal{G}_T$, $\mathsf{risk}(\mathcal{G}_T, \epsilon) \leq 1 - \delta$, if and only if $\mathcal{G}_T$ satisfies (ϵ, δ)-dissimilarity.*

Proof: [Theorem 1] (NECESSITY) If the partition $\mathcal{G}_T$ violates (ϵ, δ)-dissimilarity, i.e., $\exists G \in \mathcal{G}_T$, $\exists v \in \mathcal{SV}_G$, $|\Phi_G(v, \epsilon)| - 1 > (1 - \delta) \cdot (|G| - 1)$, then trivially, $\mathsf{risk}(\mathcal{G}_T, \epsilon) \geq$

[1]We consider the collection of SA-values in a QI-group as a multi-set, and regard each SA-value as unique.

$(|\Phi_G(v,\epsilon)|-1)/(|G|-1) > (1-\delta)$, which implies a proximity breach.

(SUFFICIENCY) If $\mathcal{G}^T$ contains a proximity breach with risk at least $(1-\delta)$, then there must exist certain $G \in \mathcal{G}_T$ and certain $v \in \mathcal{SV}_G$ which violates (ϵ,δ)-dissimilarity.

Essentially, (ϵ,δ)-dissimilarity counters general proximity breach via specifying the maximum number of ϵ-neighbors that each SA-value can have, relative to the QI-group size. It captures the impact of proximate SA-values on improving the adversary's estimation, who has a prior belief of $1/|G|$ for each SA-value in G.

Nevertheless, it is insensitive to the trivial case of small-sized QI-groups with pair-wise dissimilar SA-values. In such scenarios, despite the weak proximate QI-SA association, the small cardinality of G offers the attacker with a strong prior belief, $1/|G|$, for each SA-value. To remedy this drawback, we introduce k-anonymity [23] into our framework: by requiring every QI-group to contain at least k tuples, we upper bound this prior belief with $1/k$.

Thus, (ϵ,δ)-dissimilarity, in conjunction with k-anonymity as auxiliary, can effectively thwart association attack in terms of both exact and proximate QI-SA association. We entitle this combination $(\epsilon,\delta)^k$-dissimilarity.

D. Relevance to Principles in Literatures

It is worth emphasizing again that $(\epsilon,\delta)^k$-dissimilarity makes no specific assumption regarding the underlying data model; hence, it is effective to tackle proximity breach under most existing models. In the following, we show that most generalization principles in literatures are either in-adequate in preventing proximity breach, or essentially the special instantiations of $(\epsilon,\delta)^k$-dissimilarity under the data models which they are designed for.

Principles for categorical data: Motivated by the *homogeneity breach* wherein a majority of tuples in a QI-group share an identical SA-value, l-diversity [18] and its variant (α,k)-anonymity [25] have been proposed to ensure sufficient diversity of SA-values in every QI-group; essentially, they are both special forms of $(\epsilon,\delta)^k$-dissimilarity for data models wherein different SA-values have no sense of semantic proximity, e.g., categorical data.

Let us take (α,k)-anonymity as an example. It combines k-anonymity and l-diversity, and demands that every QI-group must contain at least k tuples, and at most α-percent of these tuples carry an identical SA-value. It is trivial to notice that (α,k)-anonymity is equivalent to $(\epsilon,\delta)^k$-dissimilarity under the setting of $\epsilon = 0$ and $1-\delta \approx \alpha$.

Principles for numeric data: For data models wherein different SA-values can be strictly ordered, e.g., one-dimensional numeric data, it qualifies as a several privacy violation if the attacker can identify the victim individual's SA-value within a short interval, even though not the exact value. To address such privacy breach, a plethora of principles have been proposed for publishing numeric sensitive data, e.g., variance control [15] and (k,e)-anonymity [30]. Specifically, variance control specifies that in every QI-group, the variance of the SA-values must be above certain threshold t; (k,e)-anonymity states that every QI-group must have at least k different SA-values, and the difference between the maximum and minimum ones must be at least e. Unfortunately, it is proved in [17] that none of these principles provide sufficient protection against proximity breaches.

The principle most relevant to $(\epsilon,\delta)^k$-dissimilarity is probably (ϵ,m)-anonymity [17]; it demands that in each QI-group G, for every SA-value x in G, at most $1/m$ of the SA-values in G belong to the interval of $[x-\epsilon, x+\epsilon]$. Clearly, (ϵ,m)-anonymity is an instantiation of $(\epsilon,\delta)^k$-dissimilarity for one-dimensional numeric data, with $1/m \approx 1-\delta$. Nevertheless, targeting a specific data model, the theoretical analysis and generalization algorithms in [17] are inapplicable for addressing general proximity breach. Moreover, since m is an integer, the users can only specify their privacy requirements in a harmonic sequence manner, i.e., $\frac{1}{2}, \frac{1}{3}, \ldots$, instead of a "stepless" continuous adjustment as supported by $(\epsilon,\delta)^k$-dissimilarity.

III. CHARACTERIZATION

In this section, we present an analytical study on the characteristics of $(\epsilon,\delta)^k$-dissimilarity. Specifically, we discuss the satisfiability of $(\epsilon,\delta)^k$-dissimilarity, and expose the implicit trade-off in setting the parameters ϵ and δ.

A. Satisfiability of (Epsilon, Delta)^k- Dissimilarity

For the given microdata table T and privacy parameters (k, ϵ, δ), the first question arises as "does there exist a partition $\mathcal{G}_T$ for T that satisfies both k-anonymity and (ϵ,δ)-dissimilarity?", i.e., the *satisfiability* of $(\epsilon,\delta)^k$-dissimilarity with respect to T. Unfortunately, in general, no efficient solution exists to answer this question unless P = NP, as shown in the next theorem.

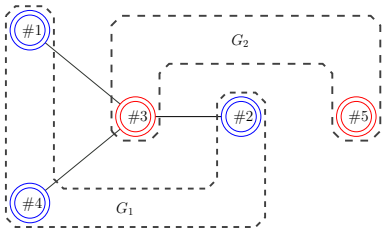

Fig. 2: Abstract graph and coloring.

Theorem 2. *In general, for the given microdata table T and parameters (ϵ, δ, k), deciding the satisfiability of $(\epsilon,\delta)^k$-dissimilarity for T is* NP*-hard.*

Before presenting the proof, we first introduce two fundamental concepts, *abstract graph* and *proper coloring*.

Definition 8 (ABSTRACT GRAPH)**.** *For a microdata table T and a proximity threshold ϵ, the abstract graph $\Psi_T^\epsilon = (\mathcal{V}_T^\epsilon, \mathcal{E}_T^\epsilon)$ is defined as follows: $\mathcal{V}_T^\epsilon$ denotes the set of vertices, with each vertex corresponding to a SA-value in T; $\mathcal{E}_T^\epsilon$ represents the set of edges over $\mathcal{V}_T^\epsilon$, and two vertices are adjacent if and only if their corresponding SA-values are ϵ-neighbors.*

Definition 9 (PROPER COLORING)**.** *Given a graph $\Psi = (\mathcal{V}, \mathcal{E})$, a (proper) m-coloring of Ψ is an assignment of no more than m colors to the vertices $\mathcal{V}$, such that no two adjacent vertices share the same color.*

Example 4. Fig. 2 illustrates the abstract graph corresponding to the SA-values appearing in the first QI-group of Table I, under the setting of $\epsilon = 0.1$. One possible 2-coloring scheme is to assign (#1, #2, #4) color 1 and (#3, #5) color 2.

Sketchily, our proof to Theorem 2 is constructed by mapping the satisfiability of $(\epsilon, \delta)^k$-dissimilarity to a proper coloring of the abstract graph corresponding to T and ϵ.

Proof: [Theorem 2] It suffices to prove that the problem under a specific setting is NP-hard. Let us consider a stringent version of $(\epsilon, \delta)^k$-dissimilarity with $\delta = 1$; that is, all SA-values in a same QI-group are required to be dissimilar.

For the given parameter ϵ and microdata T (of cardinality n), we construct an abstract graph $\Psi_T^\epsilon = (\mathcal{V}_T^\epsilon, \mathcal{E}_T^\epsilon)$. Without loss of generality, consider a partition $\mathcal{G}_T$ of T comprising m ($m \leq \lfloor n/k \rfloor$) QI-groups: $\mathcal{G}_T = \{G_i\}_{i=1}^m$. In Ψ_T^ϵ, the vertices corresponding to each G_i are assigned with one distinct color.

Clearly, under this setting, $\mathcal{G}_T$ satisfies $(\epsilon, \delta)^k$-dissimilarity, only if every two adjacent vertices in $\Psi_T(\epsilon)$ have distinct colors, i.e., a proper m-coloring. Nevertheless, it is known that determining for a general graph if a proper m-coloring exists is NP-complete [9], which implies that deciding the satisfiability of $(\epsilon, \delta)^k$-dissimilarity for given T is NP-hard.

Therefore, instead of attempting to seek the exact answer to whether an $(\epsilon, \delta)^k$-dissimilarity-satisfying partition exists for given T, we are more interested in developing approximate solutions that can 1) provide explicit and intuitive guidance for the setting of privacy parameters, and 2) efficiently find high-quality partitions of the microdata.

B. Trade-off between Epsilon and Delta

In our framework, the privacy requirement is specified as a triple of parameters k, ϵ, and δ, which however demonstrate implicit conflicts. In this section, we assume that k is fixed, and discuss the trade-off between ϵ and δ; in Section V, we reveal the quantitative relationship among ϵ, δ, and k.

Specifically, ϵ specifies the upper bound of semantic distance between two SA-values to be considered as semantically proximate. Meanwhile, δ controls the alarm threshold of proximity breach; a larger δ implies a lower tolerance of proximity-privacy risk. Clearly, by increasing ϵ or δ, one can achieve better privacy protection against association attack, though along different dimensions. An inherent trade-off, however, exists between ϵ and δ. Here, we expose this trade-off in an informal manner, and provide a quantitative modeling in Section IV.

Consider two extreme settings of ϵ. 1) $\epsilon = 0$, which amounts to saying that all distinct SA-values are considered dissimilar. By evenly distributing tuples sharing an identical SA-value into different QI-groups, a majority of the SA-values in each QI-group tend to be pairwise dissimilar; therefore, one expects to achieve high δ. 2) $\epsilon = \infty$, which means that all SA-values in the domain are considered similar. In this case, no partition can achieve any $\delta < 1$; that is, the risk of proximity-privacy breach is always 1. Intuitively, as ϵ increases, the number of pairs of similar SA-values grows, rendering it harder to achieve high dissimilarity (large δ in each QI-group), and vise versa.

IV. THEORY

As discussed in Section II, the key to anonymizing a microdata table T through generalization is to determine a partition $\mathcal{G}_T$ of T. It is however shown in Theorem 2 that deciding exactly if T is "anonymizable" for given parameters (k, ϵ, δ) is NP-hard. In this section, we establish the theoretical foundation for an approximate solution that allows intuitive and flexible tuning of the multiple privacy parameters, and finds high-quality partitions with polynomial complexity.

More concretely, we re-formulate the problem of finding an $(\epsilon, \delta)^k$-dissimilarity-satisfying partition in the framework of *defect graph coloring*; we map it to a novel *relaxed equitable coloring* problem that embeds all the privacy parameters. We then conduct an analytical study on the sufficient conditions (in terms of ϵ, δ, and k) for the existence of a valid coloring. The constructive nature of the proofs naturally leads to algorithms that are efficient in both theory and practice.

A. Problem Re-formulation

It is shown in Section III that given a microdata table[2] T and a proximity threshold ϵ, one can construct an abstract graph $\Psi^\epsilon = (\mathcal{V}^\epsilon, \mathcal{E}^\epsilon)$. A partition $\mathcal{G}$ of T corresponds to a m-coloring of Ψ^ϵ (may not be proper), which partitions the vertices $\mathcal{V}^\epsilon$ into m *color classes*, defined as below.

Definition 10 (COLOR CLASS)**.** *A m-coloring of a graph $\Psi = (\mathcal{V}, \mathcal{E})$ partitions $\mathcal{V}$ into m disjoint subsets (color classes) $\{V_i\}_{i=1}^m$, each corresponding to one distinct color.*

Next, we intend to re-formulate the problem of finding a $(\epsilon, \delta)^k$-dissimilarity-satisfying partition $\mathcal{G}$ in the framework of graph coloring. Sufficiently and necessarily, if a partition $\mathcal{G} = \{G_1\}_{i=1}^m$ of T satisfies $(\epsilon, \delta)^k$-dissimilarity, then there must exist a corresponding coloring of Ψ^ϵ that satisfies the following conditions: 1) Ψ^ϵ is colored using m colors ($\{V_i\}_{i=1}^m$ represent the m color classes); 2) the size of every color class is at least k, i.e., $|V_i| \geq k$ ($1 \leq i \leq m$); and 3) for any $v \in V_i$ ($1 \leq i \leq m$), at most $(1 - \delta) \cdot (|V_i| - 1)$ vertices in V_i are adjacent to v.

We note that the coloring problem above can be considered as a "relaxed" version of the classic proper coloring (Definition 9), in the sense that it allows a constrained number of monochromatic edges (called *defects*). It however deviates from the conventional setting of *defect coloring* as studied in graph theory, e.g., [4], in the sense that it imposes constraints on the size of every color class.

Therefore, in developing our solution, we target the following *relaxed equitable coloring* problem.

[2]Without ambiguity, in the rest of the paper, we omit this referred microdata table in the notations.

notation	definition
ϵ	threshold of semantic proximity
δ	threshold of breach
k	parameter of k-anonymity
n	cardinality of microdata table T
m	number of color classes $\lfloor n/k \rfloor$
t	$\lfloor (1-\delta)\cdot(k-1) \rfloor$
$\mathcal{G}_T^\epsilon$	abstract graph for given ϵ and T
Θ_T^ϵ	maximum degree of $\mathcal{G}_T^\epsilon$
g	lower bound of the number of movable classes

TABLE II: List of symbols and notations.

Definition 11 (RELAXED EQUITABLE $(\lfloor \frac{n}{k} \rfloor, \delta)$-COLORING)**.** *A relaxed equitable $(\lfloor \frac{n}{k} \rfloor, \delta)$-coloring of a graph Ψ^ϵ satisfies the following conditions:*

(i) Ψ^ϵ is colored using $m = \lfloor \frac{n}{k} \rfloor$ colors, with the corresponding color classes denoted by $\{V_i\}_{i=1}^m$;

(ii) the sizes of any two color classes differ by at most 1;

(iii) for any $v \in V_i$ $(1 \le i \le m)$, at most $\lfloor (1-\delta)\cdot(|V_i|-1) \rfloor$ vertices in V_i are adjacent to v.

Clearly, this coloring scheme incorporates both k-anonymity (condition (ii)) and (ϵ, δ)-dissimilarity (condition (iii)). Note that for ease of presentation, here we limit the size of every color class to be either k or $k+1$ (i.e., equitable coloring); our results however can be readily extended to support different group sizes. The details are referred to our technical report [1] due to the space constraint.

To the best of our knowledge, there is no previous study on such relaxed equitable coloring problem; therefore, the solution presented here is interesting in its own right from the perspective of graph theory.

B. Rationale

Following, we present the theoretical rationale of our equitable $(\lfloor \frac{n}{k} \rfloor, \delta)$-coloring scheme. We begin with introducing the fundamental concepts. The complete list of notations used in the presentation can be found in Table II.

Among the properties of Ψ^ϵ, we are particularly interested in its maximum degree, which is formally defined as below.

Definition 12 (MAXIMUM DEGREE)**.** *The maximum degree Θ^ϵ of a graph Ψ^ϵ is defined as $\Theta^\epsilon = \max_{v \in \mathcal{V}^\epsilon} \mathcal{D}_{\Psi^\epsilon}(v)$, where $\mathcal{D}_{\Psi^\epsilon}(v)$ denotes the degree of v in Ψ^ϵ.*

For the sake of clarity, we assume that n is divisible by k; thus, every color class is of identical size k. We use $m = \frac{n}{k}$ to denote the number of color classes, and $t = \lfloor (1-\delta)\cdot(k-1) \rfloor$ to represent the maximum number of neighbors that a vertex is allowed to have in its self-colored class.

Let $\mathcal{D}_V(v)$ denote the number of neighbors of a vertex v in a color class V. Clearly, if v has overlarge $\mathcal{D}_V(v)$ in its self-colored class V, it violates (ϵ, δ)-dissimilarity, formally

Definition 13 (VIOLATION/MOVABLE CLASS)**.** *Given a color class V and a vertex v, if $v \in V$ and $\mathcal{D}_V(v) \ge (t+1)$, v is called a violation; if $v \notin V$ and $\mathcal{D}_{V'}(v) \le t$, V is called a movable class for v, or v is movable to V.*

We have the following lemma that establishes a lower bound on the number of movable classes for any v.

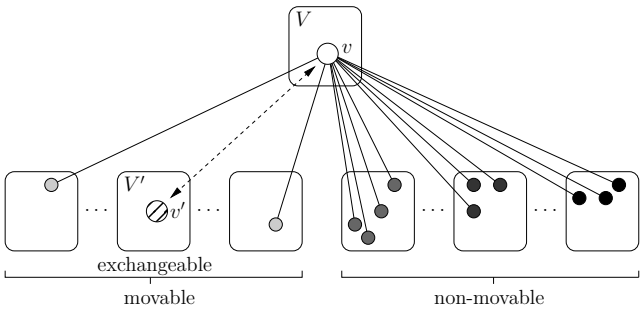

Fig. 3: Movable-classes and exchangeable classes.

Lemma 1. *Given a graph $\Psi^\epsilon = (\mathcal{V}^\epsilon, \mathcal{E}^\epsilon)$, and a m-coloring of Ψ^ϵ, $\mathcal{C} = \{V_i\}_{i=1}^m$, for any $v \in \mathcal{V}^\epsilon$, at least $g = m - \lfloor \Theta^\epsilon/(t+1) \rfloor$ color classes $V \in \mathcal{C}$ satisfy $\mathcal{D}_V(v) \le t$.*

Proof: [Lemma 1] Summing the degrees of v over all the color classes, we have $\sum_{i=1}^m \mathcal{D}_{V_i}(v) \le \Theta^\epsilon$. According to the *pigeon-hole* principle, one can derive that at most $\lfloor \Theta^\epsilon/(t+1) \rfloor$ classes contain more than t neighbors of v, from which follows this lemma.

Assume that an initial coloring $\mathcal{C} = \{V_i\}_{i=1}^m$ violates (ϵ, δ)-dissimilarity. The key idea of transforming $\mathcal{C}$ to an (ϵ, δ)-dissimilarity-satisfying coloring $\mathcal{C}' = \{V_i'\}_{i=1}^m$ is to move every violation $v \in V$ $(V \in \mathcal{C})$ to a movable class V' with $\mathcal{D}_{V'}(v) \le t$. In order to satisfy the requirement that all color classes are of identical size, it is necessary to move a vertex $v' \in V'$ back to V, i.e., v and v' are exchanged. The movement of adding v' to V, however, could potentially create more violations in V, thereby making this transformation process never converge.

To remedy this, we introduce the concept of *global potential* as an indication of the convergence of this process. Specifically, the global potential of a coloring with respect to a graph is defined as the total number of monochromatic edges. Informally, if every move of the transformation makes the potential decrease, the process will converge in polynomial time (the maximum possible potential of a coloring with respect to Ψ^ϵ is $|\mathcal{E}^\epsilon|$). Now, we proceed to formulating the impact of each move over the global potential.

Definition 14 (POTENTIAL CHANGE)**.** *The change in potential resulted from moving a vertex v from color class V to V', $\Lambda(V \xrightarrow{v} V')$, is calculated as $\Lambda(V \xrightarrow{v} V') = \mathcal{D}_{V'}(v) - \mathcal{D}_V(v)$, i.e., the change in the number of monochromatic edges.*

We demand that the move of switching two vertices $v \in V$ and $v' \in V'$ is allowed only if the global potential is decreased, formally

Definition 15 (EXCHANGEABLE CLASS)**.** *A vertex v is exchangeable to a color class V' (or V' is an exchangeable class for v) only if (i) v is movable to V', and (ii) $\exists v' \in V'$, $\Lambda(V \xrightarrow{v} V') + \Lambda(V' \cup \{v\} \xrightarrow{v'} V \setminus \{v\}) < 0$.*

This scenario is illustrated in Fig. 3: among the family of movable classes exists an exchangeable class V' that contains a vertex v' such that switching v and v' results in the decrease

of the global potential.

We are thus interested in investigating the existence of exchangeable class among the family of movable classes for v. In the following lemma, we show that if the maximum degree Θ^ϵ is bounded by certain threshold, there must exist at least one exchangeable class for v.

Lemma 2. *If $\Theta^\epsilon \leq \frac{m\cdot(t+1)}{2}$, for an arbitrary coloring $\mathcal{C} = \{V_i\}_{i=1}^m$ and any $v \in V$ with $\mathcal{D}_V(v) \geq (t+1)$, there exists at least one exchangeable class V' for v.*

Proof: [Lemma 2] Otherwise, assume that all the color classes of $\mathcal{C}$ are non-exchangeable for v. Consider the family of movable classes for v. Without loss of generality, assume that $\{V_i\}_{i=1}^g$ are movable for v. Applying the assumption of $\mathcal{D}_V(v) \geq (t+1)$, we have the following two facts:

1) $\sum_{i=1}^g \mathcal{D}_{V_i}(v) \leq \Theta^\epsilon - (m-g)\cdot(t+1)$, derived from Definition 13;

2) $\Lambda(V \xrightarrow{v} V_i) \leq \mathcal{D}_{V_i}(v) - (t+1)$ $(1 \leq i \leq g)$, derived from Definition 14.

According to the assumption, none of the movable classes are exchangeable for v; therefore, any vertex $v' \in V_i$ $(1 \leq i \leq g)$ should satisfy the next condition:

$$\Lambda(V_i \cup \{v\} \xrightarrow{v'} V \setminus \{v\}) \geq -\Lambda(V \xrightarrow{v} V_i)$$

We thus have the following inequality:

$$\begin{aligned}
\mathcal{D}_{V\setminus\{v\}}(v') &= \Lambda(V_i \cup \{v\} \xrightarrow{v'} V \setminus \{v\}) + \mathcal{D}_{V_i\cup\{v\}}(v') \\
&\geq \Lambda(V_i \cup \{v\} \xrightarrow{v'} V \setminus \{v\}) \\
&\geq -\Lambda(V \xrightarrow{v} V_i) \\
&\geq (t+1) - \mathcal{D}_{V_i}(v)
\end{aligned}$$

Summing $\mathcal{D}_{V\setminus\{v\}}(v')$ over all the vertices in the family of movable classes $\{V_i\}_{i=1}^g$ for v, we can obtain

$$\begin{aligned}
\sum_{i=1}^{g}\sum_{v'\in V_i} \mathcal{D}_{V\setminus\{v\}}(v') &\geq \sum_{i=1}^{g}\sum_{v'\in V_i} [(t+1) - \mathcal{D}_{V_i}(v)] \\
&= g\cdot k\cdot(t+1) - k\sum_{i=1}^{g}\mathcal{D}_{V_i}(v) \\
&\geq g\cdot k\cdot(t+1) \\
&\quad -k\cdot[\Theta^\epsilon - (m-g)\cdot(t+1)] \\
&= m\cdot k\cdot(t+1) - k\cdot\Theta^\epsilon \\
&\geq k\cdot\Theta^\epsilon
\end{aligned}$$

It is thus derived that the maximum degree of the vertices in $V\setminus\{v\}$ is at least $\frac{\sum_{i=1}^g\sum_{v'\in V_i}\mathcal{D}_{V\setminus\{v\}}(v')}{k-1} > \Theta^\epsilon$, which is a contradiction to the maximality of Θ^ϵ.

Based on Lemma 2, we are ready to introduce the following theorem, which can be considered as a significant extension of the classic result of Lovász [12] along the dimension of equitable coloring.

Theorem 3. *For given m, a graph $\Psi = (\mathcal{V}, \mathcal{E})$ with maximum degree Θ can be equitably colored using m colors, with each color class of degree at most $(\frac{2\Theta}{m} - 1)$, in time $O(|\mathcal{E}|\cdot|\mathcal{V}|)$.*

Proof: [Theorem 3] Start with an arbitrary initial coloring wherein all the color classes are of identical size. Consider a vertex v in a class V with more than $(\frac{2\Theta}{m} - 1)$ self-colored neighbors. As proved in Lemma 2, there must exist at least one vertex v' in an exchangeable class V' for v such that switching v and v' decreases the potential of the graph. We exchange the colors of v and v', thereby decreasing the overall number of monochromatic edges in the graph by at least 1. Repeat this process until all the violations are removed. This takes at most $|\mathcal{E}|$ steps, with the cost of each step at most $|\mathcal{V}|$, leading to the overall complexity of $O(|\mathcal{E}|\cdot|\mathcal{V}|)$.

V. XColor Algorithm

Theorem 3 paves the way for developing efficient solutions to fulfilling $(\epsilon,\delta)^k$-dissimilarity. In this section, we bridge the gap between the theoretical methodology and the practical generalization algorithm.

Concretely, we aim at addressing three key challenges. 1) The setting of the privacy parameters ϵ, δ, and k. As exposed in Section III, their inherent conflicts make it necessary to carefully balance these parameters in order to achieve the best quality of privacy protection. 2) The utility of the anonymized data. So far we have been solely focusing on providing adequate protection against proximity breach; while in real applications, it is imperative to take account of the resulted data utility when designing the anonymization algorithm. 3) The efficiency of the algorithm. Although it is difficult to further improve the asymptotic complexity of the basic algorithm in Theorem 3, one can construct a high-quality initial coloring that leads to significant performance gains over unattended initial configurations.

Following, we first reveal how to set the privacy parameters to achieve the best quality of protection, then show how to optimize the data utility along the process of anonymization, and finally discuss how to speed up the anonymization process.

A. Setting of K, Epsilon, and Delta

Within our framework, the privacy requirement is specified as a triple of parameters k, ϵ, and δ. Intuitively, k represents the lower bound of the QI-group size, thereby guaranteeing that an attacker is unable to associate a victim with a single SA-value with high confidence; ϵ indicates the threshold of semantic proximity, and two SA-values with distance below ϵ are considered as semantically similar; δ denotes the alarm threshold of proximity breach, and any breach with risk above $(1-\delta)$ needs to be eliminated.

Clearly, increasing k, ϵ, or δ all improves the quality of protection, but along different dimensions. The adjustment, however, is not arbitrary, as constrained by their inherent trade-offs. Our framework provides an explicit and quantitative modeling of such trade-offs using the following inequality (derived from Lemma 2).

$$\Theta^\epsilon \leq \frac{n}{2}\cdot[1 - (1-\frac{1}{k})\cdot\delta] \tag{4}$$

At the first glance, it may seem that the parameter ϵ is not reflected in this inequality. we note, however, that for

given microdata T, there is a surjective mapping from ϵ to the abstract graph Ψ^ϵ; that is, ϵ uniquely determines the structure of Ψ^ϵ. In particular, the density of Ψ^ϵ is directly correlated with ϵ; therefore, the maximum degree Θ^ϵ is a proper indicator of the underlying parameter ϵ: a larger ϵ implies a denser Ψ^ϵ and thus a higher maximum degree.

Different users tend to possess varying preferences for the importance of these three parameters. Below, using the case of fixed ϵ as an example, we show how to set δ and k to achieve the optimal protection. Note that stricter privacy protection is obtained at the cost of reduced data utility. Here, we concentrate our discussion on the maximum achievable protection, and temporarily ignore the utility issue.

For fixed ϵ (fixed Θ^ϵ), one can increase both δ and k when Inequality 4 holds. When it reaches the bottleneck, i.e., $(1 - 1/k) \cdot \delta = (1 - 2\Theta^\epsilon/n)$, depending on the user's preference, one can trade k (or δ) in order to increase δ (or k) further.

B. Incorporation of Data Utility

A key consideration missing in the basic algorithm derived from Theorem 3 is the utility of the resulted anonymized data. We intend to incorporate the optimization of data utility in the anonymization process.

As discussed in Section II, generalization transforms the QI-values in each QI-group to a uniform format. Evidently, this operation is performed at the cost of information loss; and the objective therefore is to minimize the global information loss in the process of anonymization.

Various metrics (a detailed survey in [11]) have been proposed to measure the information loss incurred by the generalization operation. Although our framework makes no specific assumption regarding the metrics in use, in our experiments, we employ the following model [17], [28], [29] to take the gauge of the information loss for a QI-group G:

$$\mathsf{loss}(G) = \sum_{i=1}^{d} \frac{|G.A_i^{qi}|}{|A_i^{qi}|}$$

where $|A_i^{qi}|$ represents the domain length of an QI-attribute A_i^{qi}, $|G.A_i|$ represents the length of the generalized QI-value of G, and d is the number of QI-attributes. Note that since all the QI-groups are of identical size, we omit the factor of group size in this model.

In our framework, we adopt a greedy strategy to minimize the global information loss. Specifically, when exchanging a vertex $v \in V$ with another one $v' \in V'$, the change of information loss can be formulated as follows:

$$\begin{aligned}\mathsf{loss}'(V \xleftrightarrow{v,v'} V') \quad = \quad & \mathsf{loss}(V \setminus \{v\} \cup \{v'\}) \\ & + \mathsf{loss}(V' \setminus \{v'\} \cup \{v\}) \\ & - \mathsf{loss}(V) - \mathsf{loss}(V')\end{aligned}$$

We seek the pair of vertices $(\mathring{v}, \mathring{v}')$ for exchange that lead to the maximum decrease in the overall information loss, formally

$$(\mathring{v}, \mathring{v}') = \arg\min_{v,v'} \mathsf{loss}'(V \xleftrightarrow{v,v'} V')$$

At the same time instance, however, a significant number of vertices might all violate (ϵ, δ)-dissimilarity, and each might correspond to a considerable number of exchangeable vertices, resulting in the prohibitive complexity of finding the optimal pair $(\mathring{v}, \mathring{v}')$. In our implementation, we make further approximation by dividing this operation into two steps: in the first step, we find the vertex $\mathring{v}$ (in a class V) whose departure minimizes the information loss of V, formally

$$\mathring{v} = \arg\min_{v} \mathsf{loss}(V \setminus \{v\}) - \mathsf{loss}(V)$$

In the second step, we seek the vertex $\mathring{v}'$ (in a class V') whose exchange with $\mathring{v}$ minimizes the overall information loss of V and V', formally

$$\begin{aligned}\mathring{v}' = \arg\min_{v'} \quad & \mathsf{loss}(V \setminus \{\mathring{v}\} \cup \{v'\}) + \mathsf{loss}(V' \cup \{\mathring{v}\} \setminus \{v'\}) \\ & - \mathsf{loss}(V) - \mathsf{loss}(V')\end{aligned}$$

Assuming that at the time instance there are n_v instances of v, each corresponding to $n_{v'}$ instances of v', the approximation here reduces the complexity from $O(n_v \cdot n_{v'})$ to $O(n_v + n_{v'})$.

C. Optimization of Initial Coloring

As proved in Theorem 3, for an arbitrary initial configuration, the basic algorithm can converge to an equitable (m, δ)-coloring with time complexity of $O(|\mathcal{V}| \cdot |\mathcal{E}|)$, where the term $|\mathcal{E}|$ follows the upper bound of the number of monochromatic edges. We deem it as the key of decreasing the computational complexity to minimize the initial number of monochromatic edges. The following lemma provides the rationale for our initial configuration construction procedure.

Lemma 3. *Assume that the probability that two vertices are adjacent is proportional to the product of their degrees. The overall number of monochromatic edges is minimized if all color classes have identical sum of degrees.*

Proof: [Lemma 3] Let $\{d_i\}_{i=1}^n$ denote the degrees of the vertices $\{v_i\}_{i=1}^n$, correspondingly, and $\mathsf{edge}(v_i, v_j)$ be the probability that two vertices v_i and v_j share an edge. According to the assumption, we have $\mathsf{edge}(v_i, v_j) \propto d_i \cdot d_j$.

We intend to partition the vertices into m color classes of identical size k, $V_c = \{v_i^c\}_{i=1}^k$ $(1 \le c \le m)$, such that the overall number of monochromatic edges is minimized:

$$\min \sum_{c=1}^{m} \sum_{v_i^c, v_j^c \in V_c} \mathsf{edge}(v_i^c, v_j^c)$$

Given the assumption that $\mathsf{edge}(v_i^c, v_j^c) \propto d_i^c \cdot d_j^c$, one can derive that this problem is equivalent to minimizing the following simplified version.

$$\min \sum_{c=1}^{m} \left(\sum_{i=1}^{k} d_i^c \right)^2 - \sum_{i=1}^{n} d_i^2$$

For fixed $\{d_i\}_{i=1}^n$, we can omit $\sum_{i=1}^n d_i^2$. Now, let X_c denote $\sum_{i=1}^k d_i^c$. We have the following equivalent formulation.

$$\min \sum_{c=1}^{m} X_c^2 \quad s.t. \quad \sum_{c=1}^{m} X_c = \sum_{i=1}^{n} d_i$$

It is well known that under this setting the minimum is achieved when $X_i = X_j$ for every pair of i and j.

Algorithm 1: INITIALIZE (Ψ, k)

Input: graph $\Psi = (\mathcal{V}, \mathcal{E})$, k
Output: m color classes ($m = \lfloor n/k \rfloor$, $n = |\mathcal{V}|$)

1 sort $\{v | v \in \mathcal{V}\}$ in descending order of degrees as $v_1, v_2, \ldots, v_n$ (with degrees $d_1, d_2, \ldots, d_n$, respectively);
2 initialize m empty buckets $\{V_j\}_{j=1}^{m}$;
3 **for** i *from* 1 *to* $k \cdot m$ **do**
4 $j^* \leftarrow \arg\min_j d(V_j)$ s.t. $h(V_j) < k$;
5 add v_i to V_{j^*};
  `// process the remaining vertices`
6 **for** i *from* $(k \cdot m + 1)$ *to* n **do**
7 $j^* \leftarrow \arg\min_j d(V_j)$ s.t. $h(V_j) < k + 1$;
8 add v_i to V_{j^*};
9 return $\{V_j\}_{j=1}^{m}$

Unfortunately, reductible from the *equal-subset-sum* problem [26], constructing an initial configuration with all the color classes of identical sum of degrees is NP-hard.

Following, we present a heuristic solution to constructing the initial coloring, which is empirically proved to lead to fairly small number of monochromatic edges. Algorithm 1 outlines the INITIALIZE procedure. It takes as input a graph Ψ (of cardinality n) and a parameter k, and generates a coloring of Ψ with m color classes ($m = \lfloor n/k \rfloor$). For each class V_j, it keeps track of its sum of degrees $d(V_j)$ and its height $h(V_j)$ (the number of vertices assigned to V_j). INITIALIZE balances $d(V_j)$ $(1 \leq j \leq m)$ in a greedy manner: at each iteration, from the pool of unassigned vertices, it picks the one with the maximum degree, and adds it to a non-full class with the minimum sum of degrees (lines 3-5). After all the classes are filled with k vertices, it assigns remaining vertices (if n is not divisible by k) to classes with the minimum sum of degrees (line 6-8).

D. A Complete Framework

Incorporating the suite of multi-folded optimizations into the basic algorithm derived from Theorem 3, we are now ready to present the complete anonymization framework, XCOLOR, as sketched in Algorithm 2. Given the microdata table T and the parameters ϵ, δ, and k, XCOLOR produces an anonymized table T^* that satisfies $(\epsilon, \delta)^k$-dissimilarity.

Specifically, for the given T and ϵ, XCOLOR first constructs an abstract graph Ψ_T^ϵ (line 2), and invokes INITIALIZE to obtain an initial coloring of Ψ_T^ϵ (line 3). It then identifies and eliminates all the violations of (ϵ, δ)-dissimilarity (line 4-11): at each iteration, XCOLOR switches a violation $\mathring{v}$ (in class V) with an exchangeable vertex $\mathring{v}'$ (in class V'), meanwhile minimizing the information loss. Finally, the coloring is mapped to a partition of the microdata table T (line 12); and by generalizing all the corresponding QI-groups, an anonymized table T^* is obtained (line 13).

Algorithm 2: XCOLOR (T, k, ϵ, δ)

Input: microdata table T, parameters k, ϵ, δ
Output: anonymized table T^*

1 $n = |T|$, $m = \lfloor n/k \rfloor$, $t = \lfloor (1 - \delta) \cdot (k - 1) \rfloor$;
2 construct an abstract graph $\Psi_T^\epsilon = (\mathcal{V}_T^\epsilon, \mathcal{E}_T^\epsilon)$;
  `// create an initial coloring`
3 $\{V_i\}_{i=1}^{m}$ = INITIALIZE (Ψ_T^ϵ, k);
  `// vertices exchanges`
4 $\mathcal{V} = \{v | v \in V_i$ and $\mathcal{D}_{V_i}(v) \geq t + 1$ $(1 \leq i \leq m)\}$;
5 **while** $\mathcal{V} \neq \emptyset$ **do**
6 find $\mathring{v} = \min_{v \in \mathcal{V}} \mathsf{loss}(V \setminus \{v\}) - \mathsf{loss}(V)$;
7 $\mathcal{V}' \leftarrow$ exchangeable vertices for $\mathring{v}$;
8 find $\mathring{v}' = \min_{v' \in \mathcal{V}'} \mathsf{loss}(V \setminus \{\mathring{v}\} \cup \{v'\})$
9 $+\mathsf{loss}(V' \cup \{\mathring{v}\} \setminus \{v'\}) - \mathsf{loss}(V) - \mathsf{loss}(V')$;
10 switch $\mathring{v}$ and $\mathring{v}'$;
11 update $\mathcal{V}$;
  `// map to partition`
12 map $\{V_i\}_{i=1}^{m}$ to a partition $\mathcal{G}_T = \{G_i\}_{i=1}^{m}$;
13 generalize G_i $(1 \leq i \leq m)$ and return T^*;

VI. EXPERIMENTS

In this section, we perform an empirical evaluation to validate the analytical models and the efficacy of the proposed countermeasure. The experiments comprise two main parts: 1) we intend to study the impact of general proximity breach over the anonymized data generated according to alternative privacy definitions; 2) we aim to investigate the practical performance of the XCOLOR method, in terms of proximity-privacy protection, utility preservation and operation efficiency.

A. Experimental Setting

Our experiments use a real dataset SAL (http://ipums.org), which has now become a de facto benchmark for evaluating anonymized data publishing algorithms, e.g., [17], [28], [29]. The dataset contains 50k valid tuples, each corresponding to the personal information of an American adult, collected from the US census. The attributes used in the experiments, their domain lengths (the number of distinct values), and the heights of their domain generalization taxonomies (for categorical attributes) are listed in Table III.

attribute	type	d. length	d. height
Age	numeric	85	N/A
Sex	categoric	2	2
Marital Status	categoric	6	2
Race	categoric	9	2
Education	sensitive	17	N/A
Work Class	sensitive	10	4
Income	sensitive	50	N/A

TABLE III: Attributes of the SAL dataset.

We use the first four attributes as QI-attributes, and regard the last three as a composite sensitive attribute. The semantic proximity between two sensitive values is defined as their normalized L^p space distance (within the interval [0,1]); the weight of each component in the attribute is set according to a health report [24] on the influence of education, income, and occupation to cardiovascular disease.

All the algorithms are implemented in C++, and the experiments are conducted on a Linux workstation running 1.7GHz Pentium III and 2GB memory.

B. Experimental Results

Attack Vulnerability: In the first set of experiments, we intend to evaluate the impact of general proximity breach over the "publishable" data generated according to alternative privacy definitions. Specifically, we deploy an implementation of Mondrian [14], a state-of-the-art anonymizaton algorithm, to generate l-diverse tables. To show that even strong l-diversity does not guarantee sufficient protection against general proximity breach, we fix the value of l to be 20, which implies that in every QI-group, no more than 5% tuples can share identical sensitive values.

We then measure the risk of proximity breach (Definition 6) for the published data by counting the number of QI-groups which violate (ϵ, δ)-dissimilarity. Given a generalized table, we define its vulnerability to association attack as the proportion of QI-groups that contain proximity breach over the total number of QI-groups in the table. Moreover, we measure the vulnerability under varying settings of semantic proximity metric, i.e., L^1- and L^2-norm.

Fig. 4 gives a contour view of the vulnerability (with respect to ϵ and δ) of the published table generated by Mondrian with l = 20. The left and right plots correspond to L^1- and L^2-norm, respectively. It is clear that in both cases the vulnerability increases significantly as ϵ or δ grows. For example, in the left plot (L^1-norm), for any $\epsilon \geq 0.18$ and $\delta \geq 0.85$, over half of the QI-groups are subjected to general proximity breach. One can also notice the "mutual enforcement" of the two parameters ϵ and δ with respect to the vulnerability: for a large ϵ (or δ), a trivial increase in δ (or ϵ) leads to a considerable growth of the attack vulnerability.

Because of the similar characteristics of L^p-norms ($p = 1, 2$ in the example above), in the following experiments, we primarily use L^1-norm as the semantic proximity metric.

Data Quality: We measure the utility of the generalized data using the approach described in [17], [28], [29]. Consider the *count queries* of the form

```
select count (*) from generalized-data
where A_1 ∈ I_1 and A_2 ∈ I_2 and ... and A_q ∈ I_q
```

where $A_1, \ldots, A_q$ are q distinct random attributes, comprising qd QI-attributes and qs SA-attributes (qd + qs = q), and each I_i ($1 \leq i \leq q$) is a randomly selected interval in the domain of A_i. The length of the interval is controlled by a parameter s ($0 \leq s \leq 1$), called the *query selectivity*. Specifically, the length $|I_i|$ of I_i is given by $|I_i| = |A_i| \cdot s^{\frac{1}{q}}$, where $|A_i|$ is the domain length of A_i. Clearly, a larger s implies a wider selection interval. In our experiments, qs is fixed to 2, and qd is called the *query dimensionality*.

The quality of the resulted data is measured by the average relative error of answering such count queries using the generalized table. Specifically, for each query, we evaluate it over the generalized table using the approach described in [14], and calculate the relative error as $|gen - raw|/raw$, where gen and raw represent the evaluation results over the generalized table and the microdata table, respectively.

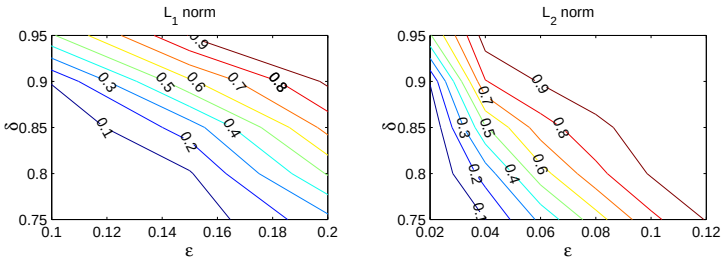

Fig. 4: Vulnerability of the published table (generated by Mondrian with l = 20) to general proximity breach.

In each set of experiments, we fix three of the parameters, ϵ, δ, k, and s, and evaluate the impact of the rest one over the quality of the generalized table. Each set of experiments contain a workload of 1k queries. The default setting of the parameters is: ϵ = 0.1, δ = 0.8, k = 10, and s = 0.1.

From left to right, Fig. 5 plots the data quality (measured by average relative error) with respect to ϵ, δ, k, and s. First notice that the data utility is a decreasing function of both ϵ and δ, though demonstrating fairly different patterns. This is expected, since a larger ϵ or δ indicates stricter privacy requirement, at the cost of reduced data utility. Meanwhile, the sharp increase of the relative error for δ from 0.8 to 0.9 is contributed to the fact that a majority of (ϵ, δ)-dissimilarity violations fall in this interval; for the given ϵ and initial configuration, the data utility is usually reversely correlated with the number of operations needed to remedy these violations. Nevertheless, note that in both cases, the relative error is always below 15%, even in the extreme case of query dimensionality $qd = 3$.

The influence of the parameter k over the data utility shows more interesting patterns. The relative error decreases as k varies from 5 to 15, reaches its minimum at k = 15, and then slightly grows afterwards. Here is an intuitive explanation: for fixed ϵ and δ, we claim that smaller k usually results in a larger number of (ϵ, δ)-dissimilarity violations. To see this, consider the following simple example, which can be readily generalized to support our claim:

Example 5. Assume a random initial partition, i.e., a tuple is assigned to every QI-group with identical probability, a total number of 8 tuples, and δ = 1/2. Consider a tuple x with two ϵ-neighbors x_1 and x_2 for given ϵ. If k = 4, x causes a violation if x_1 and x_2 are assigned to the same QI-group as x, i.e., with probability $(1/2)^2 = 1/4$; if k = 2, x results in a violation if either x_1 or x_2 is assigned to x's group, thus with a larger probability 1-$(3/4)^2$ = $7/16$.

As we have pointed out, the data utility is usually reversely correlated with the number of operations needed to remedy the violations. This explains the high relative error for a small k. Meanwhile, recall that the QI-attribute values in a QI-group is generalized to their minimum bounding interval (for quantitative attribute) or their lowest common ancestor on the hierarchy (for categorical attribute), a larger k implies more significant distortions, which explains the high relative error for a large k.

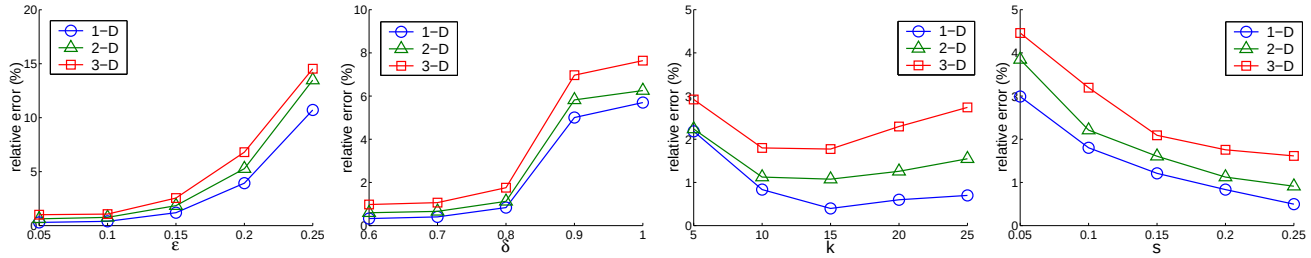

Fig. 5: Average relative error with respect to four parameters ϵ, δ, k, and s.

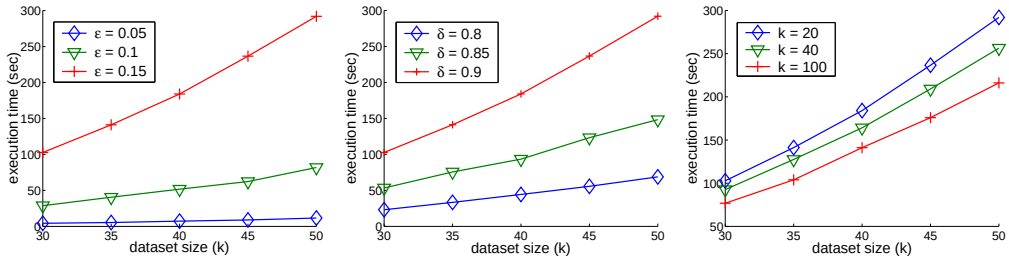

Fig. 6: Average execution time with respect to ϵ, δ, k, and the size of the dataset.

The rightmost plot of Fig. 5 shows the data utility as a function of the query selectivity s, which varies from 0.05 to 0.25. It is noticed that the average workload error decreases as s grows, which is consistent with the analysis in [14]: the generalized data enjoys higher accuracy for queries with larger selection intervals. We note that even for s as low as 0.05, our approach guarantees high data utility, with the relative error below 5%.

Execution Efficiency: Fig. 6 plots the average execution time of our generalization algorithm with respect to the dataset size which varies from 30k to 50k. Furthermore, in each set of experiments, we fix two of the three parameters ϵ, δ, and k, and measure the impact of the remaining one on the running time. By default, the parameter setting is: ϵ = 0.15, δ = 0.9, and k = 20.

One can notice that in all three plots, the average execution time grows approximately linearly with respect to the dataset size, and all the experiments terminate within minutes. As expected, the computation cost appears as an increasing function of ϵ and δ. This is contributed to that for a given initial partition, a larger ϵ or δ results in a more significant number of (ϵ, δ)-dissimilarity violations, and the computational cost is usually proportional to the number of operations needed to remedy these breaches. It is also interesting to notice that in the rightmost plot, the cost decreases as k grows, which empirically validates our analysis regarding the impact of k over the number of (ϵ, δ)-dissimilarity violations.

VII. Related Work

The problem of centralized publication has attracted intensive research recently. The existing literatures can be classified mainly into two categories. The first one aims at devising privacy definitions and principles, as the criteria to measure the quality of the protection provided by an anonymization model. As mentioned in our discussion, k-anonymity [23], l-diversity [18], and (α, k)-anonymity [25] target association attack based on exact QI-SA association, while t-closeness [16], (k, e)-anonymity [30], and (ϵ, m)-anonymity [17] take into consideration the attack leveraging proximate association. For scenarios with different background knowledge assumptions, a collection of principles have been proposed: δ-presence [21] assumes that the adversary has no prior knowledge regarding the presence of individuals in the microdata; (c, k)-safety [19] and privacy skyline [5] consider the case that the adversary possesses external knowledge regarding the target individual, other individuals, or the family of individuals sharing a same SA-value; m-invariance [29] is designed for sequential releases of microdata; while differential privacy [6] measures the quality of protection from the perspective of the perturbation mechanism itself.

The second category of work explores the possibility of fulfilling the proposed anonymization principles, meanwhile preserving the data utility to the maximum extent. In [2], Aggarwal showed that due to the curse of dimensionality, it is hard to enforce even 2-anonymity for high-dimensional microdata. In [2], [14], [20], it was proved that finding the optimal k-anonymization with minimum information loss is NP-Hard for the suppression, multi-dimensional, and attribute models, respectively. Despite these negative results, it was shown in [3], [13] that the optimal relation can be found efficiently by systematically enumerating the possible generalizations, in conjunction of effective pruning using heuristics. Efficient greedy-manner solutions have also be considered [8], [14], [27]. Besides the heuristic methods above, a set of approximation algorithms have been developed [2], [20], [22], which provide theoretical guarantees on the quality of the resulting data. Another direction of work attempts to optimize the data utility, without compromising the hard privacy requirement. Instead of anonymizing the whole microdata table,

Kifer and Gehrke [11] advocated anonymizing and publishing a set of marginals, to ameliorate the curse of dimensionality. In [28], [30], it was proposed to publish the QI-attributes and SA-attributes separately, so as to preserve the utility of the QI-values. LeFevre et al. [15] differentiated the generalization level for different subsets of the microdata according to their importance, and propose a workload-aware anonymization scheme.

Graph coloring has been a prominent topic in graph theory for a long history. An (ordinary vertex) coloring is a partition of the vertices of a graph into independent sets. It is known that determining if a general graph can be colored with less than k colors (its chromatic number) is NP-Hard [9]. Many variants and generalizations have been considered, particularly in relation to practical applications. Cowen et al. [4] considered a relaxation of coloring in which the color classes partition the vertices into subgraphs of degree at most d, called (k, d)-coloring, following the classic work of Lovász [12]. In [7], Erdös considered the problem of equitable coloring, imposing the constraint that each color class should be of identical size, and made the famous conjecture that the chromatic number of a graph with maximum degree Θ is at most $(\Theta+1)$, which was later proved in [10]. However, to the best of our knowledge, no previous work exists on the problem of equitable coloring with defect, as discussed in this paper.

VIII. Conclusion and Future Work

This work represents a systematic study on the problem of protecting general proximity privacy, with findings of broad applicability. Our contributions are multi-folded: we highlighted and formalized proximity breach in a data-model-neutral manner; we proposed a unified privacy definition, $(\epsilon, \delta)^k$-dissimilarity, with theoretically guaranteed protection against association attack in terms of both exact and proximate QI-SA association; we derived the criteria that enable to efficiently check the satisfiability of the principle for given microdata; we developed a novel anonymization model, XColor, to fulfill this principle, which offer flexible control over multiple privacy requirements; we conducted extensive experiments over real data to verify the practical efficacy of the XColor method.

This work also opens several directions for future research. First, in this paper, we define proximity breach as a set of SA-values proximate to a common one in a QI group (star structure). When other topological structures are taken into consideration, e.g., clique, chain, and cycle, how to measure and remedy the possible privacy breach is an important problem. Second, in developing our solution, we only allow the vertices to be exchanged between two color classes. While enabling vertices to be transferred in a cycle, e.g., moving vertex v_1 from class V_1 to V_2, v_2 from V_2 to V_3, and v_3 from V_3 to V_1, the solution is still valid, and we envision a better bound than that provided in this paper. Third, in our generalization framework, we apply a suite of heuristics to optimize the statistical utility of the resulting data (a best-effort strategy). It is worth investigating how to incorporate utility as a first-class citizen in designing the anonymization solution.

References

[1] "Protecting general proximity privacy". Technical Report.

[2] C. Aggarwal. "On k-anonymity and the curse of dimensionality". In *VLDB*, 2005.

[3] R. Bayardo and R. Agrawal. "Data privacy through optimal k-anonymity". In *ICDE*, 2005.

[4] L. Cowen, W. Goddard and C. Jesurum. "Coloring with defect". In *SODA*, 1997.

[5] B. Chen, R. Ramakrishnan and K. LeFevre. "Privacy skyline: privacy with multidimensional adversial knowledge". In *VLDB*, 2007.

[6] C. Dwork. "Differential privacy". In *ICALP*, 2006.

[7] P. Erdös. "Some applications of probability of graph theory and combinatorial problems". Theory of Graphs and its Applications, Publ. House Czechoslovak Acad. Sci., Prague, 1964.

[8] B. Fung, K. Wang and P. Yu. "Top-down specialization for informaiton and privacy preservation". In *ICDE*, 2005.

[9] M. Garey and D. Johnson. "Computers and intractability: a guide to the theory of NP-completeness". Freeman, San Francisco, CA, 1981.

[10] A. Hajnal and E. Szemerédi. "Proof of a conjecture of P. Erdös". Combinatorial theory and its application, II. North-Holland, Amsterdam, 1970.

[11] D. Kifer and J. Gehrke. "Injecting utility into anonymization databases". In *SIGMOD*, 2006.

[12] L. Lovász. "On decompositions of graphs". *Studia Sci. Math. Hungar.*, 1:237-238, 1966.

[13] K. LeFevre, D. DeWitt and R. Ramakrishnan. "Incognito: efficient full-domain k-anonymity". In *SIGMOD*, 2005.

[14] K. LeFevre, D. DeWitt and R. Ramakrishnan. "Mondrian multidimensional k-anonymity". In *ICDE*, 2006.

[15] K. LeFevre, D. DeWitt and R. Ramakrishnan. "Workload-aware anonymization". In *SIGKDD*, 2006.

[16] N. Li, T. Li and S. Venkatasubramanian. "t-closeness: privacy beyond k-anonymity and l-diversity". In *ICDE*, 2007.

[17] J. Li, Y. Tao and X. Xiao. "Preservation of proximity privacy in publishing numerical sensitive data". In *SIGMOD*, 2008.

[18] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. "l-diversity: privacy beyond k-anonymity". In *ACM TKDD*, 1(1), 2007.

[19] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke and J. Halpern. "Worst-case background knowledge in privacy". In *ICDE*, 2007.

[20] A. Meyerson and R. Williams. "On the complexity of optimal k-anonymity". In *PODS*, 2004.

[21] M. Nergiz, M. Atzori and C. Clifton. "Hiding the presence of individuals from shared databases". In *SIGMOD*, 2007.

[22] H. Park and K. Shim. "Approximate algorithm for k-anonymity". In *SIGMOD*, 2007.

[23] L. Sweeney. "k-anonymity: a model for protecting privacy". In *International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems*, 10(5), 2002.

[24] M. Winkleby, D. Jatulis, E. Frank and S. Fortmann. "Socioeconomic status and health: how education, income, and occupation contributes to risk factors for cardiovascular disease". *Am J Public Health*, 82(6), 1992.

[25] R. Wong, J. Li, A. Fu and K. Wang. "(alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing". In *SIGKDD*, 2006.

[26] G. Woeginger and Z. Yu. "On the equal-subset-sum problem". *Information Processing Letter*, 42, 1992.

[27] K. Wang, P. Yu and S. Chakraborty. "Bottom-up generalization: a data mining solution to privacy protection". In *ICDM*, 2004.

[28] X. Xiao and Y. Tao. "Anatomy: Simple and effective privacy preservation". In *VLDB*, 2006.

[29] X. Xiao and Y. Tao. "m-invariance: towards privacy preserving re-publication of dynamic datasets". In *SIGMOD*, 2007.

[30] Q. Zhang, N. Koudas, D. Srivastava and T. Yu. "Aggregate query answering on anonymized tables". In *ICDE*, 2007.