Privacy Risk in Anonymized Heterogeneous Information Networks

Aston Zhang*
University of Illinois at Urbana-Champaign
lzhang74@illinois.edu

Xing Xie
Microsoft Research
xingx@microsoft.com

Kevin Chen-Chuan Chang
University of Illinois at Urbana-Champaign
kcchang@illinois.edu

Carl A. Gunter
University of Illinois at Urbana-Champaign
cgunter@illinois.edu

Jiawei Han
University of Illinois at Urbana-Champaign
hanj@illinois.edu

XiaoFeng Wang
Indiana University at Bloomington
xw7@indiana.edu

ABSTRACT

Anonymized user datasets are often released for research or industry applications. As an example, *t.qq.com* released its anonymized users' profile, social interaction, and recommendation log data in KDD Cup 2012 to call for recommendation algorithms. Since the entities (users and so on) and edges (links among entities) are of multiple types, the released social network is a *heterogeneous information network*. Prior work has shown how privacy can be compromised in homogeneous information networks by the use of specific types of graph patterns. We show how the extra information derived from heterogeneity can be used to relax these assumptions. To characterize and demonstrate this added threat, we formally define privacy risk in an anonymized heterogeneous information network to identify the vulnerability in the possible way such data are released, and further present a new de-anonymization attack that exploits the vulnerability. Our attack successfully de-anonymized most individuals involved in the data—for an anonymized 1,000-user *t.qq.com* network of density 0.01, the attack precision is over 90% with a 2.3-million-user auxiliary network.

Keywords

privacy, anonymization, attack, social networks, data mining

1. INTRODUCTION

The world is getting more inter-connected. Tons of social network data are generated through people's interactions, and different entities are linked across multiple relations, forming a gigantic information-rich, inter-related and multi-typed *heterogeneous information network* [5]. Is there any risk in the current efforts to avoid privacy intrusion upon the anonymized copy of a heterogeneous information network? We start with a motivating example.

*Part of the work was completed at Microsoft Research.

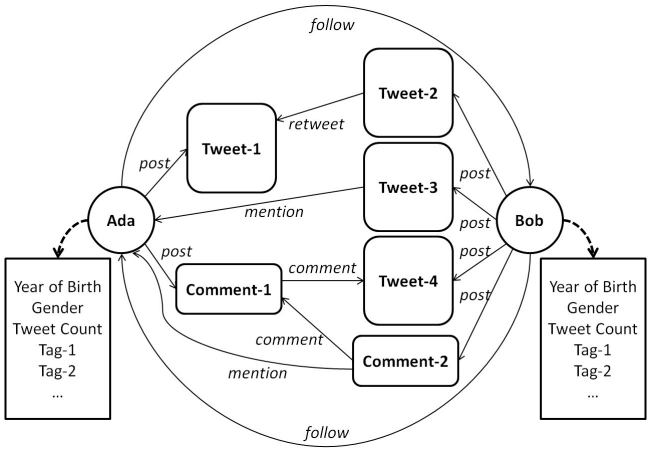


Figure 1: The heterogeneous information network in t.qq

1.1 Motivating Example

Various datasets containing *micro-data*, that is, information about specific individuals, have been released for different research purposes or industry applications [11]. Some datasets contain individual profiles, preferences, or transactions, which many people consider sensitive or *private*. In the recent KDD Cup 2012, *t.qq.com* (a popular microblogging site, hereinafter referred to as *t.qq*) released its 2.3 million users' profile, social interaction, and recommendation preference log data to call for more efficient recommendation algorithms [1]. In a microblogging site like *t.qq* as depicted in Figure 1, entities (nodes) correspond to *users*, *tweets* or *comments*, and edges correspond to different types of links (*post*, *mention*, *retweet*, *comment*, and *follow*) among them[1]. Since both nodes and links are of multiple types, such a social network is essentially a heterogeneous information network [14]. Besides identifying information such as *user ID* which has been anonymized by randomly assigned strings, some other attributes are also replaced with meaningless IDs, such as *user tags*.

[1]The terms *edge* and *link* are used interchangeably in this work, while the term *entity* is preferred over *node* here to reflect more realistic scenarios where each node contains multiple attributes rather than a single identifier in the settings of a heterogeneous information network.

10.5441/002/edbt.2014.53

In the released anonymized *target dataset*, consider an adversary that is interested in breaching privacy of some selected target users based on their preferences. The preference can be inferred from the target users' recommendation preference (acceptance/rejection) log included in the target dataset. This information is sensitive and not accessible on the *t.qq* site (the rejection log cannot be inferred from the site). Suppose the adversary obtains the non-anonymized *auxiliary dataset* from *t.qq* exactly containing the users from the same time-synchronized target dataset. To *de-anonymize* the users of interests in the target dataset, the adversary has to match the meaningless user IDs in the target dataset with the real user names in the auxiliary dataset. Given the rich information available in the heterogeneous information network as demonstrated in Figure 1, suppose the adversary locks his target on an anonymized user (say, *A3H*) in the target dataset who accepted the "follow Citibank" recommendation but rejected all other bank recommendations. The adversary may search in the auxiliary dataset by specifying *A3H*'s entity profile (*A3H*'s year of birth, hereinafter referred to as yob: 1980, gender: male, *etc.*) combined with *A3H*'s multiple social links (mention, retweet, comment, follow) and profile information of its neighbor entity to whom the target user connects via these links—*A3H* gave 15 comments to an anonymized female user *F8P* born in 1985 and retweeted an anonymized male user *M7R* 10 times that is born in 1970. If Ada in the non-anonymized auxiliary dataset is the only one that satisfies the matching—Ada has both the same profile information as *A3H* and Ada has the same social interactions with the other users of the same gender and age as those of *F8P* and *M7R* correspondingly; thus, the adversary successfully de-anonymizes *A3H* by establishing a *unique matching* between it in the target dataset and the real user Ada in the auxiliary dataset. Now the adversary knows Ada probably has a Citibank account or is interested in applying for it. The leak of such private information may allow scammers to spam Ada with phishing URLs camouflaged with the Citibank online-banking interface. In fact, 8% of some sampled 25 million URLs posted to microblogging sites point to phishing, malware, and scams [4].

Therefore, there is *privacy risk* in an anonymized heterogeneous information network if such unique matchings can be easily established. Users in a network of high privacy risk that can be easily de-anonymized may be vulnerable to external threats. In this work, we experimentally substantiate adversaries can exploit the privacy risk to de-anonymize over 90% users in a 1,000-user *t.qq* network of density 0.01 from a 2,320,895-user auxiliary network.

1.2 Limitations of k-Anonymity

To formalize privacy risk observed in Section 1.1, directly using the existing metric seems possible at first thought. A dataset is said to be k-anonymous if on the minimal set of attributes in the table that can be joined with external information to de-anonymize individual records, each record is indistinguishable from at least $k-1$ other records within the same dataset [16]. The larger the value of k, the better the privacy is preserved.

Consider target dataset T_{1000} that satisfies 1000-anonymity and another target dataset T_2 that satisfies 2-anonymity, together with their original non-anonymized counterparts. Imagine a new tuple t^* is created and inserted into both T_{1000} and T_2. After anonymization processes still no any other tuple in either dataset has the same value of t^*, and the new datasets are T^*_{1000} and T^*_2 respectively. Both T^*_{1000} and T^*_2 are now 1-anonymity simply because of the injection of t^*—both T^*_{1000} and T^*_2 are same vulnerable in terms of the same k-anonymity. Suppose a selective adversary is not interested in de-anonymizing t^*, then the remaining T^*_{1000} of 1000-anonymity seems much less vulnerable than the remaining T^*_2 of 2-anonymity, which may be misled by the same 1-anonymity.

Due to limitations of k-anonymity in differentiating individuals in the same target dataset, it is not suitable to formalize privacy risk in a more general scenario where adversaries may not be equally interested in de-anonymizing all users. In this paper we define privacy risk in a more general sense, and prove it can be very high in the anonymized heterogeneous information network.

1.3 New Settings, New Threats

Social media are getting popular with more and more functionalities. As shown in Section 1.1, *t.qq* allows its over 500 million users to connect with one another in different ways such as follow, mention, retweet, and comment. The growing multi-typed heterogeneous information networks out of the growing social media functionalities may render the existing homogeneous information network anonymization algorithms no more effective.

Existing de-anonymization attacks on social networks made several assumptions, such as both target and auxiliary graphs are large-scale so random graphs or non-trivial cliques can be re-identified from both graphs [2, 12]. It should be highlighted that, in the new settings of a heterogeneous information network, if new attacks are feasible while relaxing these assumptions, such attacks must be addressed in the proposal of all relevant anonymization algorithms.

1.4 Our Contributions

In this work we make three unique contributions. First, we propose a definition of privacy risk tuned to the concerns of heterogenous information networks. In particular, this definition considers a more general situation where adversaries may not be equally interested in compromising all users' privacy. We show that the privacy risk can be high in an anonymized heterogeneous information network, and can be exploited in practice.

Second, we present a de-anonymization algorithm against heterogeneous information networks which exploits the identified privacy risk without requiring creating new accounts or relying on easily-detectable graph structures in a large-scale network. While central in illuminating the privacy issue for a heterogeneous information network, we also expect our algorithm to be applied to de-anonymizing a homogeneous information network (with slight performance degradation).

Our third contribution is a practical evaluation of the KDD Cup 2012 *t.qq* anonymized dataset, which contains 2.3 million users and over 60 million multiple types of social links among them. To demonstrate the effectiveness of the de-anonymization algorithm, we apply the state-of-the-art graph anonymization algorithms to the *t.qq* dataset, which were claimed effective by their designers for defending graph structural attacks. The experiments show that our algorithm is able to beat the investigated graph anonymization algorithms in the settings of a heterogeneous information network even without knowledge of the specific anonymization technique in use. It undermines the notion of "security by obscurity" for privacy preservation: ignorance of the anonymization does not prevent an adversary from de-anonymizing successfully.

2. RELATED WORK

Simply replacing sensitive information with random strings cannot guarantee privacy and how to release data for different research purposes or industry applications without leaking any privacy information has been an interesting problem.

2.1 Relational Data Anonymization

A major category of privacy attacks on relational data is to de-anonymize individuals by joining a released table containing sen-

sitive information with some external tables modeling the auxiliary dataset of attackers. To mitigate this type of attacks, k-anonymity was proposed [16]. Further enhanced techniques include l-diversity [9] and t-closeness [7].

Narayanan and Shmatikov proposed de-anonymization attacks against high-dimensional micro-data and showed success in Netflix Prize dataset [11]. They pointed out micro-data are characterized by high dimensionality and sparsity. A recent study by Narayanan *et al.* further demonstrated the feasibility of internet-scale author identification via linguistic stylometry [10]. However, all the aforementioned studies assume that an adversary utilizes *attribute information* of micro-data and can deal with relational data only.

2.2 Graph Structural Attacks

In a large-scale social network, it is hard to observe non-trivial random subgraphs or cliques [13]. Hence they easily stand out if they exist. Backstrom *et al.* discussed active attacks where adversaries create users and establish connections randomly among them and attach such random subgraphs ("sybil nodes") into the target nodes in the auxiliary graph data [2]. Since such random subgraphs can be easily detected from the anonymized counterpart of the original data, the target nodes connected to the sybil nodes are then de-anonymized by consulting the original auxiliary graph. Narayanan and Shmatikov pointed out the main drawback of this active attack is that, creating accounts, links among themselves and links to target nodes, is not feasible on a large-scale [12]. They designed an attack propagating the de-anonymization process via neighbor structure from the initial precisely-matched "seed nodes". Hence success of this attack heavily depends on if such seed nodes can be detected precisely; thus, seed nodes must stand out easily both in the target and auxiliary dataset. So non-trivial cliques are chosen [12]. Since there is no guarantee that the released anonymized network is always large, this attack is not always successful because non-trivial cliques cannot always be detectable.

2.3 Graph Data Anonymization

For graph-based social network data, the degree of nodes in a graph can reveal the identities of individuals. Liu and Terzi studied a specific graph-anonymization problem and called a graph k-degree anonymous if for every node v, there exist at least $k-1$ other nodes in the graph with the same degree [8]. This definition of anonymity prevents de-anonymization of individuals by adversaries with a background knowledge of the degree of certain nodes.

Zhou and Pei identified a structural *neighborhood attack* and tackled it by proposing k-neighborhood anonymization [19]. They assumed an adversary may know the neighbors of the target nodes and their inter-connections. The privacy preservation goal is to protect neighborhood attacks which use neighbor structure matching to de-anonymize nodes. For a social network, suppose an adversary knows the neighbor structure for a node. If such neighbor structure has at least k isomorphic copies in the anonymized social network, then the node can be de-anonymized in the target dataset with confidence at most $1/k$ [20]. Due to its heavy isomorphism testing computation, a limitation of this attack is only distance-1 neighbors can be evaluated effectively.

Zou *et at.* assumed an attacking model where an adversary can know any subgraph that contains the targeted individual and proposed k-automorphic anonymity that the graph must has $k-1$ non-trivial automorphism and no node is mapped to itself under the $k-1$ non-trivial automorphism [21]. Wu *et al.* proposed a similar k-symmetry [18].

Cheng *et al.* identified that k-automorphism approach is insufficient for protecting link privacy and proposed the k-security

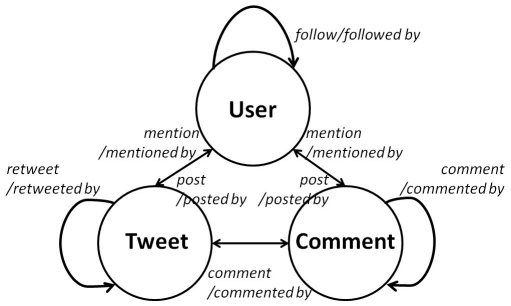


Figure 2: The corresponding network schema for the heterogeneous information network in Figure 1

anonymity [3]. In their approach, an anonymized graph satisfies k-security if for any two target individuals and any subgraphs containing either individual, the adversary cannot determine either whether a node that is linked to either target individual (NodeInfo Security) or whether both target individuals are linked by a path of a certain length (LinkInfo Security), with probability higher than $1/k$.

Although these recent graph data anonymization algorithms can be applied to social network data against graph structural attacks in Section 2.2, their applicability has not been demonstrated in the more challenging settings of a heterogeneous information network. Our evaluation in Section 6 shows that these graph data anonymization algorithms are not effective to preserve privacy of an anonymized heterogeneous information network.

3. HETEROGENEOUS INFORMATION NETWORK SETTINGS

In this section, we formalize the general anonymized heterogeneous information network settings that are frequently discussed in the remaining of the paper and illustrate them with the motivating example discussed in Section 1.1.

DEFINITION 1. *The **information network** is a directed graph $G=(V,E)$ with an entity type mapping function $\tau: V \to \mathcal{E}$ and a link type mapping function $\phi: E \to \mathcal{L}$, where each entity $v \in V$ belongs to one particular entity type $\tau(v) \in \mathcal{E}$, and each edge $e \in E$ belongs to a particular link type $\phi(e) \in \mathcal{L}$. If two edges belong to the same link type, they must share the same starting and ending entity types.*

DEFINITION 2. *The **heterogeneous information network** is an information network where $|\mathcal{E}|>1$ or $|\mathcal{L}|>1$.*

A sample heterogeneous information network for the *t.qq* dataset is depicted in Figure 1. Given a complicated heterogeneous information network, it is necessary to provide its meta level (*i.e.*, schema-level) description for better understanding the network, and *network schema* is to describe the meta structure of a network.

DEFINITION 3. *The **network schema**, denoted as $T_G=(\mathcal{E},\mathcal{L})$, is a meta template for a heterogeneous information network $G=(V,E)$ with the entity type mapping $\tau: V \to \mathcal{E}$ and the link mapping $\phi: E \to \mathcal{L}$, which is a directed graph defined over entity types $\mathcal{E}$, with edges as links from $\mathcal{L}$.*

Figure 2 shows the network schema for the heterogeneous information network in Figure 1. In practice data publishers may not release information about all the entities and links in the original

network schema while links among the same entity type (also the target entity type of adversaries' interests) are generally available either directly or indirectly via summarization over different entity types [1]. In view of this, although we believe providing richer information about multiple types of entities could further facilitate de-anonymization, in this work, we consider a more challenging and practical scenario where data publishers only provide limited information about how the same type of entity (*i.e.*, *target entity type* $\mathcal{E}^*$) can be linked via different types of links or over different types of entities. Thus, a simplified network schema is needed such that it reflects only the relationships over the target entity type.

DEFINITION 4. *The **target meta paths (target network schema links)** $\mathcal{P}(\mathcal{E}^*)$, are paths defined on the graph of network schema $T_G = (\mathcal{E}, \mathcal{L})$, denoted by $\mathcal{E}^* \xrightarrow{\mathcal{L}_1} \mathcal{E}_1 \xrightarrow{\mathcal{L}_2} ... \xrightarrow{\mathcal{L}_n} \mathcal{E}^*$.*

DEFINITION 5. *The **target network schema** $T^*_G = (\mathcal{E}^*, \mathcal{L}^*)$ is projected from $T_G = (\mathcal{E}, \mathcal{L})$ where $\mathcal{L}^*$ are* reproduced *or* short-circuited *from target meta paths $\mathcal{P}(\mathcal{E}^*)$ and target entity type $\mathcal{E}^*$.*

To illustrate, we take the released target *t.qq* dataset as an example. This anonymized dataset contains the following files and attributes (anonymized attributes are marked with underlines):

- *recommendation preference data*: user ID($\mathcal{A}$), recommended item ID($\mathcal{R}$), result (whether $\mathcal{A}$ likes $\mathcal{R}$)
- *user profile data*: user ID, yob, gender, tweet count (no. of tweets), tag IDs
- *user mention data*: user ID($\mathcal{A}$), user ID($\mathcal{B}$), the number of times $\mathcal{A}$ mentioned $\mathcal{B}$ either in $\mathcal{A}$'s tweets or comments (mention strength)
- *user retweet data*: user ID($\mathcal{A}$), user ID($\mathcal{B}$), the number of times $\mathcal{A}$ retweeted $\mathcal{B}$'s tweets (retweet strength)
- *user comment data*: user ID($\mathcal{A}$), user ID($\mathcal{B}$), the number of times $\mathcal{A}$ commented $\mathcal{B}$ either in $\mathcal{B}$'s tweets or comments (comment strength)
- *user follow data*: user ID(follower), user ID(followee)

In the above dataset, besides user entities' profile information, users' multiple social interactions are also available. Thus, the adversary can decide to project the original network schema in Figure 2 to only reflect relationships among his target user entity. Navigating the original network schema based on the above user mention, retweet, comment, and follow data, these target meta paths connecting users across different types of entities are possible:

- *user mention path*: $User \xrightarrow{post} Tweet \xrightarrow{mention} User$ or $User \xrightarrow{post} Comment \xrightarrow{mention} User$ (short-circuited feature: mention strength)
- *user retweet path*: $User \xrightarrow{post} Tweet \xrightarrow{retweet} Tweet \xrightarrow{posted\ by} User$ (short-circuited feature: retweet strength)
- *user comment path*: $User \xrightarrow{post} Comment \xrightarrow{comment} Tweet \xrightarrow{posted\ by} User$ or $User \xrightarrow{post} Comment \xrightarrow{comment} Comment \xrightarrow{posted\ by} User$ (short-circuited feature: comment strength)
- *user follow path*: $User \xrightarrow{follow} User$

The target meta paths allow the adversary to produce a new network schema by projecting the original network schema to a simplified one to only reflect particular few relationships over the target entity type. Specifically, the user mention, retweet and comment paths can be *short-circuited* to produce new links over users respectively while the user following path can be *reproduced* in the projection. It is also emphasized that, the target meta paths are

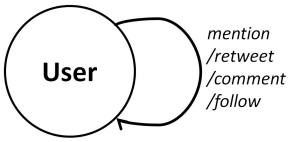


Figure 3: The target network schema for Figure 2

able to greatly enrich the features (attributes) of the target entity by utilizing different *distances* of *neighbors* from the target entity along the specified meta paths. Specifically, target meta paths that are short-circuited across different types of entities and different types of links, may preserve the link heterogeneity information of the network by generating new *short-circuited feature (attribute)* and further enrich the features of the target entity. For instance, the short-circuited feature *mention strength* can be newly generated from the user mention path.

The target network schema for Figure 2 is shown in Figure 3. Since target meta paths may span across multiple types of entities, entity heterogeneity information is still preserved, although not fully, in target network schema only containing the target entity type.

Therefore, the de-anonymization problem in the settings of a heterogeneous information network can be formulated as follows. Detailed illustrations are provided in Section 5.

DEFINITION 6. *The **de-anonymization problem in heterogeneous information network** is utilizing the background knowledge of the public graph $G = (V, E)$, the private graph $G' = (V', E')$, and the target network schema T^*_G to de-anonymize a target entity $v' \in V'$ by establishing matches between v' and a candidate set $C \subseteq V$ where the anonymized v''s counterpart $v \in C$. If $|C| = 1$ and the only element $v \in C$ is the correct counterpart of v', the de-anonymization is successful.*

4. PRIVACY RISK ANALYSIS

Intuitively, privacy risk in a heterogeneous information network is the ease of formulating unique attribute-metapath-combined values as formalized in Section 3. Formal analysis is derived from the definition of privacy risk in general anonymized datasets.

4.1 Attribute-Metapath-Combined Values of Target Entities

Data publishers anonymize data through generalization, suppression, adding, deleting, switching edges or nodes [15][20]. Naturally, such modifications cause information loss and for a certain privacy preservation goal they should be minimized to ensure the anonymized data still satisfy the need for how they are expected to be used, *i.e.*, the need for *utility*. Generally, a certain level of utility has to be preserved for the anonymized *t.qq* dataset in order to design effective and reliable recommendation algorithms; thus, an adversary is expected to be able to compromise some sacrificed privacy due to the natural tradeoff between utility and privacy preservation [20]. In the *t.qq* dataset case, the utility is preserved in the sense that, some attribute values of user entities and most of the social interactions among different user entities are preserved (non-anonymized) as in the available target dataset descriptions in Section 3 (*e.g.*, non-anonymized attributes are not underlined).

Based on the target network schema in Figure 3, Figure 4 describes an example of how user entities are directly inter-connected via part of different types of links in the *t.qq* dataset. Here m, r, c, f stands for *mention*, *retweet*, *comment*, *follow* links in the target network schema shown in Figure 3.

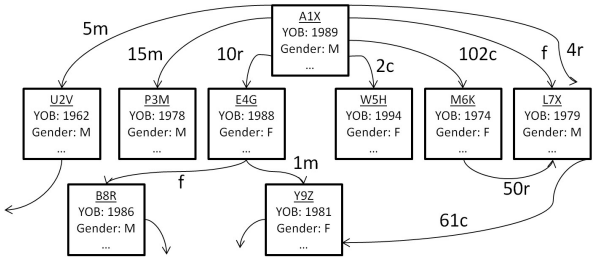


Figure 4: The neighbors of the target entity *A1X* are generated along target meta paths

As mentioned in Section 3, target meta paths that are short-circuited across different types of entities and different types of links preserve the link heterogeneity information of the information network and further enrich the features of the target entity. It should be noted that, following the user mention path identified in Section 3, $5m$ in Figure 4 from *A1X* to *U2V* indicates a new numerical feature (attribute) short-circuited from the user mention path—the mention strength from *A1X* to *U2V* in the target dataset of value 5 either through the tweet entity or comment entity. Thus, multiple meta-paths inject richer heterogeneity information for target entities in the settings of a heterogeneity information network.

If target user entities in the target dataset can form unique *attribute-metapath-combined values* across the entire network, these users can be de-anonymized from the auxiliary dataset by establishing unique matches and the dataset is not secure. To analyze the privacy risk of a heterogeneous information network, which can be intuitively considered similar to the ease of formulating unique attribute-metapath-combined values, one way is to expand the attribute dimensions of micro-data by navigating from user entities to their neighbors, neighbors' neighbors, and so on, via their multiple types of target meta paths.

With the assumption made in Section 1.1 that the target and auxiliary datasets are time-synchronized counterparts, take *A1X* in Figure 4 as an example. Without utilizing meta paths and only utilizing profile attribute information, the features of *A1X* are:

- Max. Distance-0: *yob, gender, ...*

After utilizing his immediate distance-1 neighbors along target meta paths, the features of *A1X* are expanded to (here "5-time-mentionee" means a mentionee mentioned 5 times by the target entity, *i.e.*, mention strength = 5):

- Max. Distance-1: *yob, gender, ..., 5-time-mentionee (U2V)'s yob, 5-time-mentionee's gender, ..., 15-time-mentionee (P3M)'s yob, 15-time-mentionee's gender, ..., 10-time-retweetee (E4G)'s yob, 10-time-retweetee's gender, ...*

Further utilizing his distance-2 neighbors (neighbors of distance-2 along target meta paths from *A1X*), the features of *A1X* are further expanded to:

- Max. Distance-2: *yob, gender, ..., 5-time-mentionee's yob, 5-time-mentionee's gender, ..., 15-time-mentionee's yob, 15-time-mentionee's gender, ..., 10-time-retweetee's yob, 10-time-retweetee's gender, ..., 10-time-retweetee's followee (B8R)'s yob, 10-time-retweetee's followee's gender, 10-time-retweetee's 1-time-mentionee (Y9Z)'s yob, 10-time-retweetee's 1-time-mentionee's gender, ...*

Consistent with the idea by Narayanan and Shmatikov that large dimensions of micro-data give rise to risks of privacy [11], the expansion of dimensions by propagating via multiple types of target meta paths seems to increase the possibility for a user entity to form a unique attribute-metapath-combined value under all the expanded features across the entire dataset, which can be considered as privacy risk. In the remaining of this section, we formally prove this intuition from the observations.

4.2 Privacy Risk in General Anonymized Datasets

Privacy Risk indicates risk that privacy of a given dataset can be compromised—the higher privacy risk, the lower security and *vice versa*. Hence it might be tempting to directly adopt the notion of widely-used k-anonymity and simply reverse its value to obtain the measure of privacy risk. Here we state that, k-anonymity is not able to differentiate users from one another in terms of their different levels of security or privacy risk.

As discussed in Section 1.2, k-anonymity may be misleading in more general situations where adversaries may not be equally interested in compromising all users' privacy. To address its limitations, when quantifying risk of any user in any dataset, we consider factors that influence privacy risk both socially and mathematically.

In real life, it is highly possible that an adversary is not equally interested in compromising everyone's privacy in a dataset. As illustrated in Section 1.1, an adversary may be more motivated to de-anonymize an anonymized user who probably has a Citibank account. We denote the loss function of tuple t_i by $l(t_i)$, with values between 0 and 1. $l(t_i)$ can be considered as the potential loss of a user whose privacy is compromised given that this user does care about his loss of privacy. Therefore, in a social network, $l(t_i)$ is a certain user's privacy need because such need is positively correlated with the cost of privacy breach; hence, it is the *social factor* of a user's privacy risk.

Similar to the concept of k-anonymity, we make the same assumption that the target dataset is an anonymized copy of the same auxiliary dataset. In any given dataset T, if there are $k(t_i) - 1$ other tuples of the same value of tuple t_i, the probability that each of these $k(t_i)$ tuples, say t_i, can be de-anonymized by random guessing with probability no higher than $\frac{1}{k(t_i)}$. Therefore, the higher value of $\frac{1}{k(t_i)}$, the higher possibility that the privacy of user t_i can be compromised—hence the higher privacy risk of the user t_i. $\frac{1}{k(t_i)}$ is the *mathematical factor*. Mathematical factor can be considered positively correlated with the attack incentive as well: given the same social factor, the adversary is more motivated to de-anonymize the user with a higher mathematical factor because the potential attack precision is higher.

Combining both social and mathematical factors, we define the privacy risk of a tuple in a dataset as follows.

DEFINITION 7. *We define the **privacy risk** $\Re(t_i)$ of tuple t_i in dataset T as follows:*

$$\Re(t_i) = \frac{l(t_i)}{k(t_i)},$$

where $k(t_i)$ is the number of tuples in T with the same value of tuple t_i, and $l(t_i)$ is the loss function of tuple t_i.

Averaging the risk $\Re(t_i)$ for each tuple t_i in dataset T, the risk $\Re(T)$ for dataset T is defined as follows.

DEFINITION 8. *The **privacy risk** $\Re(T)$ of dataset T is*

$$\Re(T) = \frac{\sum_{i=1}^{N} \Re(t_i)}{N},$$

where size N is the number of tuples t_i in T.

It is noted that the privacy risk value $\mathfrak{R}(T) \in [0, 1]$. Denoting by $\mathbb{C}(T)$ the *cardinality* of T—the number of distinct values, or distinct combined values under different attributes, describing each tuple t_i in T, we give the following lemma.

LEMMA 1. *Given dataset T with the cardinality $\mathbb{C}(T)$, for each tuple t_i in T, assuming the loss function is independent of $\frac{1}{k(t_i)}$ with mean value μ, the expected privacy risk*

$$\mathbb{E}(\mathfrak{R}(T)) = \frac{\mu\mathbb{C}(T)}{N}.$$

PROOF. By Definition 7 and 8,

$$\mathfrak{R}(T) = \frac{\sum_{i=1}^{N} \frac{l(t_i)}{k(t_i)}}{N}.$$

$$\begin{aligned}
\mathbb{E}(\mathfrak{R}(T)) &= \frac{\sum_{i=1}^{N} \mathbb{E}(\frac{1}{k(t_i)})\mathbb{E}(l(t_i))}{N} \\
&= \frac{\sum_{i=1}^{N} \mu\mathbb{E}(\frac{1}{k(t_i)})}{N} \\
&= \frac{\mu\mathbb{E}(\sum_{i=1}^{N} \frac{1}{k(t_i)})}{N} \\
&= \frac{\mu\mathbb{E}(\mathbb{C}(T))}{N} \\
&= \frac{\mu\mathbb{C}(T)}{N}.
\end{aligned}$$

□

Lemma 1 provides an estimation of dataset privacy risk in a relatively general sense. For instance, if the loss function for each tuple is a random number between 0 and 1 and independent of $\frac{1}{k(t_i)}$, the expected privacy risk of the dataset is $\frac{\mathbb{C}(T)}{2N}$. Although it may be interesting to quantify the social factor in other ways, to guarantee the highest possible privacy need from all users has been considered, in the remaining analysis we focus on the mathematical factor and set the value of every loss function $l(t_i)$ to 1. Adversaries may still have varying attack incentives in terms of different mathematical factors as discussed earlier in this section.

THEOREM 1. *The privacy risk $\mathfrak{R}(T)$ of dataset T is*

$$\mathfrak{R}(T) = \frac{\mathbb{C}(T)}{N}, \quad (\mathfrak{R}(T) \in [\frac{1}{N}, 1]),$$

where in T, N is the number of tuples, and cardinality $\mathbb{C}(T)$ is the number of distinct (combined) attribute values describing tuples.

PROOF. The proof can be completed by applying Lemma 1 and mathematical derivation with $l(t_i) = 1$. $\mathfrak{R}(T)$ is lowest when all the tuples are of the same value; in contrast, if every t_i has a unique value in T, $\mathfrak{R}(T) = 1$. □

Back to the example of T_{1000} and T_2 in Section 1.2, suppose they are both of the same size 1000—T_{1000} has 1000 tuples of the same value while T_2 has 500 same-value tuple pairs and values from different pairs are distinct. By Definition 8, $\mathfrak{R}(T_{1000}) = 0.001$ and $\mathfrak{R}(T_2) = 0.5$ and the result is consistent with k-anonymity in terms of relative privacy risk. After inserting the unique tuple t^*, $\mathfrak{R}(T^*_{1000}) = \frac{2}{1001}$ and $\mathfrak{R}(T^*_2) = \frac{501}{1001}$, reasonably indicating T^*_{1000} is in general still much less vulnerable than T^*_2. It addresses the identified limitations of k-anonymity when adversaries may not select some users to de-anonymize in the target dataset.

4.3 Privacy Risk in Anonymized Heterogeneous Information Networks

Section 4.1 informally shows entity attribute dimensions grow fast when neighbors are utilized. It is highlighted that, rather than the exact value of privacy risk, it is the growth of privacy risk with respect to max. distances of utilized neighbors n that we focus on. Hence, given any anonymized dataset, the number of tuples N is fixed as a constant. So Theorem 1 implies that privacy risk $\mathfrak{R}(T)$ is of the same order of growth as that of the cardinality $\mathbb{C}(T)$.

THEOREM 2. *For power-law distribution of the user out-degree, the lower and upper bounds for the expected heterogeneous information network cardinality grows faster than double exponentially with respect to the max. distance of utilized neighbors.*

PROOF. Given a network schema $T^*_G = (\mathcal{E}^*, \mathcal{L}^*)$ projected from its original schema $T_G = (\mathcal{E}, \mathcal{L})$ and the network entity size N is ideally large enough and all possible distinct values describing $\mathcal{E}^*$ appear in T^*_G. Let $\mathcal{A}(\mathcal{E}^*)_j$ and $\mathcal{A}(L^*_i)_j$ denote the j-th attribute of the entity type $\mathcal{E}^*$ and the link type L^*_i. We assume independence among entity attributes and link types with attributes along target meta paths. To focus on the analysis of key factors that may affect the bounds of network cardinality, we also assume an entity has at most in-degree 1, the link among each pair of entities is of all types and the out-degree k of each entity follows the power-law distribution $P_K(k) = ck^{-\alpha}$, which are commonly adopted in social network analysis with $\alpha \in [2, 3]$ [13][19].

To analyze the number of distinct attribute-metapath-combined values describing $\mathcal{E}^*$, or the cardinality $\mathbb{C}(T^*_G)$, of the network schema T^*_G, we begin with the network cardinality $\mathbb{C}(T^*_G)$ without utilizing any neighbors (distance-0); it is equal to the *entity cardinality* $\mathbb{C}(\mathcal{E}^*)$, which is the actual observed number of distinct combined attribute values describing entities:

$$\mathbb{C}(T^*_G)_0 = \mathbb{C}(\mathcal{E}^*).$$

Theoretically, $\mathbb{C}(\mathcal{E}^*)$ can be as high as the product of each entity attribute's cardinality:

$$\mathbb{C}(\mathcal{E}^*) \leq \prod_{j=1}^{|\mathcal{A}(\mathcal{E}^*)|} \mathbb{C}(\mathcal{A}(\mathcal{E}^*)_j).$$

After utilizing the distance-1 neighbors from the entity, let $\mathbb{C}(L^*_i)$ denote the *homogeneous link cardinality*, which is the actual observed number of distinct combined attribute values describing the link L^*_i. Likewise, the maximum value of L^*_i is the product of each attribute cardinality of the link type L^*_i:

$$\mathbb{C}(L^*_i) \leq \prod_{j=1}^{|\mathcal{A}(L^*_i)|} \mathbb{C}(\mathcal{A}(L^*_i)_j).$$

Since entities are connected to one another via different target meta paths, *heterogeneous link cardinality* is no greater than the product of each homogeneous link cardinality:

$$\mathbb{C}(\mathcal{L}^*) \leq \prod_{i=1}^{|\mathcal{L}^*|} \mathbb{C}(L^*_i).$$

Thus, the number of distinct values that an entity can have when distance-1 neighbors are utilized is:

$$\mathbb{C}(T^*_G)_1 = \mathbb{C}(T^*_G)_0 \cdot (\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*))^k.$$

By utilizing neighbors of next distance iteratively, generally when max. distance of utilized neighbors from target entities $n > 0$,

$$\mathbb{C}(T^*_G)_n = \mathbb{C}(T^*_G)_{n-1} \cdot (\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{k^n}. \tag{1}$$

Based on the distribution function of power law for the out-degree $P_K(k) = ck^{-\alpha}$, we estimate the expected value $\mathbb{E}[\mathbb{C}(T^*_G)_n]$ of Equation 1 as follows:

$$\begin{aligned}
\mathbb{E}[\mathbb{C}(T^*_G)_n] &= \mathbb{C}(T^*_G)_{n-1} \cdot \mathbb{E}[(\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{k^n}] \\
&\geq \mathbb{C}(\mathcal{E}^*) \cdot \mathbb{E}[(\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{k^n}] \\
&= \mathbb{E}[\mathbb{C}(\mathcal{E}^*) \cdot (\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{k^n}] \\
&= \textstyle\sum_{k=1}^{N} P_K(k) \cdot \mathbb{C}(\mathcal{E}^*) \cdot (\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{k^n} \\
&> \textstyle\sum_{k=2}^{N} ck^{-\alpha} \cdot \mathbb{C}(\mathcal{E}^*) \cdot (\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{k^n} \\
&\geq \textstyle\sum_{k=2}^{N} ck^{-\alpha} \cdot (\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{k^n}.
\end{aligned}$$

Let $f = ck^{-\alpha} \cdot (\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{k^n}, \quad k \in \mathbb{R}, \ \ 2 \leq k \leq N,$

$$\begin{aligned} \frac{\partial f}{\partial k} &= \frac{c(\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{k^n}(nk^n ln(\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n) - \alpha)}{k^{\alpha+1}} \\ &> 0 \qquad (nk^n ln(\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n) > \alpha). \end{aligned}$$

Hence,

$$\mathbb{E}[\mathbb{C}(T_G^*)_n] \ > \ 2^{-\alpha}(N-1)c \cdot (\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{2^n}.$$

Since the vertex size N is given, the lower bound of the expected network cardinality is

$$\Omega\{\mathbb{E}[\mathbb{C}(T_G^*)_n]\} = (\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{2^n}. \tag{2}$$

To establish the upper bound of the expected network cardinality, since $k \leq N$ and we assume N is large, solving the recursion of Equation 1 we have

$$\begin{aligned} \mathbb{C}(T_G^*)_n &\leq \mathbb{C}(\mathcal{E}^*)^{\frac{N^{n+1}-1}{N-1}} \mathbb{C}(\mathcal{L}^*)^{\frac{N^{n+1}((N-1)n+1)-N}{(N-1)^2}} \\ &\approx (\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{N^n}. \end{aligned}$$

Hence the upper bound of the expected network cardinality is the same as that of the network cardinality when all k is set to N:

$$O\{\mathbb{E}[\mathbb{C}(T_G^*)_n]\} = (\mathbb{C}(\mathcal{E}^*)\mathbb{C}(\mathcal{L}^*)^n)^{N^n}. \tag{3}$$

Equation 2 and Equation 3 complete the proof. □

Recalling the positive linear relationship between privacy risk and cardinality from Theorem 1, we obtain the following corollary.

COROLLARY 1. *For power-law distribution of the user out-degree, the lower and upper bounds for the expected privacy risk of a heterogeneous information network grows faster than double exponentially with respect to the max. distance of utilized neighbors.*

Corollary 1 substantiates the privacy risk growth in a heterogeneous information network as observed in Section 4.1. It should be emphasized that, it is the heterogeneity of information network links, which is in the mathematical form of $\mathbb{C}(\mathcal{L}^*)^n$, that makes both bounds even a higher order than double exponential growth.

4.4 Limitations of the Analysis

While it may be tempting to conclude that, as long as the max. distance of utilized neighbors grows infinitely, the dimensions for each entity will grow more than double exponentially until the privacy risk $\mathfrak{R}(t)$ becomes 1; it should be pointed out that it is not feasible in practice.

First, the assumption that N is large and all possible distinct values describing $\mathcal{E}^*$ appear in T_G^* may not hold. Then the observed cardinality depends on how to sample from a pool of all possible distinct values. The extreme case is that such "sampling" is so biased that each entity is assigned a value from a very small subset of the pool. However, such a "sampling" bias hardly happens because both $\mathbb{C}(\mathcal{E}^*)$ and $\mathbb{C}(\mathcal{L}^*)$ are actual observed cardinalities which are generally of reasonable sizes in practice.

Second, the assumption that in-degree is at most 1 may not hold and a large-scale information network in practice often has small average diameters [17]. For instance, in Figure 5, if user v_1' and user v_2' have the same attribute-metapath-combination value after utilizing their distance-1 neighbors, further utilizing their longer-distance neighbors will not make them unique from each other since they will share the same neighbors of distances longer than 1. In addition, the existence of leaf nodes which do not have outgoing edges also prevents utilizing longer-distance of entity neighbors,

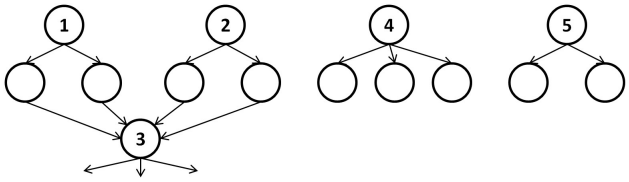


Figure 5: The bottleneck scenarios

such as user v_4' and v_5' in Figure 5. However, in Section 6 we show in practice this concern can be addressed because a slight increase of n renders the actual cardinality very close to N.

We show the empirical findings in Table 1 and Figure 7 that $\mathfrak{R}(t)$ grows very fast when $n \in \{0, 1\}$ and after $n > 1$, $\mathfrak{R}(t)$ grows towards 1 asymptotically until the bottleneck scenarios keep $\mathfrak{R}(t)$ from growing. Nonetheless, the growth order of bounds is consistent with the actual growth during $n \in \{0, 1\}$ so $\mathfrak{R}(t)$ can soon get very close to 1.

4.5 Practical Implications to Reduce Privacy Risk

To reduce privacy risk, following the two bounds established in Equation 2 and Equation 3, either the entity cardinality $\mathbb{C}(\mathcal{E}^*)$ or link cardinality $\mathbb{C}(\mathcal{L}^*)$ has to be reduced. Since preventing users from sharing their profile information may restrain the growth of online communities, practical efforts should focus on reducing $\mathbb{C}(\mathcal{L}^*)$ which makes both bounds grow more than double exponentially. Instead of making heterogenous types of links fully accessible from the public, online forums may only allow premium users to access all or partial types of relationships, so $\mathbb{C}(\mathcal{L}^*)$ decreases.

5. DE-ANONYMIZATION ALGORITHM

To exploit the privacy risk in a heterogeneous information network as identified in Section 4, a de-anonymization algorithm is presented with a threat model.

5.1 Threat Model

In the privacy risk analysis, we assume the auxiliary dataset is exactly the non-anonymized counterpart of the target dataset. Although this assumption may hold in real attack scenarios, we consider a more challenging scenario where there is a time gap between the time data publishers release the target dataset and the time adversaries start to collect the auxiliary dataset from the web. Since a social network generally grows over time, we assume the later collected auxiliary dataset contain all the target users and links among them. Other or newly formed users and links can be included in the auxiliary dataset as well.

We emphasize that de-anonymizing with the auxiliary dataset larger than the target dataset is a non-trivial and more challenging task than both datasets are of the same size, especially when allowing certain attribute values and links to grow. First, when the auxiliary dataset becomes a superset of the target dataset without increasing the cardinality of each tuple from the target dataset, the actual risk should be lower because each tuple t_i in the target dataset has potentially more matches with users in the auxiliary dataset. Second, allowing certain attribute or link growth gives rise to potentially more candidate users in the auxiliary dataset that may match a certain target user. For instance, for a user in the target dataset that posted 3 tweets and only followed 5 users, any user in the auxiliary dataset with more than 3 tweets and more than 5 followees could be a candidate match if we consider number of tweets

Algorithm 1: De-anonymizing entity v' in a Heterogeneous Information Network: DeHIN (G, G', T^*_G, v', n)

Input: $G = (V, E)$: auxiliary graph, $G' = (V', E')$: target graph, $T^*_G = (\mathcal{E}^*, \mathcal{L}^*)$: target network schema, $v' \in G'$: target entity, n: max. distance of utilized neighbors

Output: C: candidate set from the auxiliary dataset matching v'

```
begin
    C <-set- ∅;
    foreach v ∈ V
        if entity_attribute_match(v', v, E*)
            if n > 0
                if link_match(n, v', v, G, G', T*_G)
                    C <-add- v;
            else
                C <-add- v;
    return C;
```

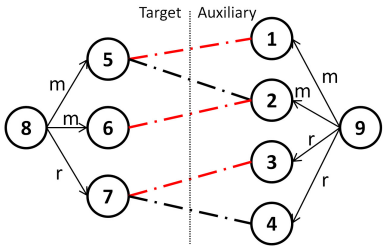


Figure 6: Comparing neighbors via multiple types of target network schema links from target and auxiliary datasets

Algorithm 2: Comparing neighbors of entities v' and v via heterogeneous links: $link\_match(n, v', v, G, G', T^*_G)$

Input: n: max. distance of utilized neighbors, $v' \in G'$: target entity, v: the entity in auxiliary graph under comparison, $G = (V, E)$: auxiliary graph, $G' = (V', E')$: target graph, $T^*_G = (\mathcal{E}^*, \mathcal{L}^*)$: target network schema

Output: is_match: a boolean value

```
begin
    is_match <-set- true;
    G_B <-set- ∅ (The bipartite graph modeling neighborhood matching);
    N_b(v', L*_i) <-set- v''s neighbors via the link type L*_i;
    N_b(v, L*_i) <-set- v's neighbors via the link type L*_i;
    foreach link type L*_i ∈ L*
        foreach neighbor b'_i ∈ N_b(v', L*_i)
            ∅ ← C(b'_i); (C(b'_i): candidate set for b'_i);
            foreach neighbor b_i ∈ N_b(v, L*_i)
                if link_attribute_match(b'_i, b_i)
                    if entity_attribute_match(b'_i, b_i)
                        if n = 1
                            C(b'_i) <-add- b_i;
                        else
                            if link_match(n − 1, v', v, G, G', T*_G)
                                C(b'_i) <-add- b_i;
            G_B <-add- C(b'_i);
        if max_bipartite_match(G_B) ≠ |N_b(v', L*_i)|
            is_match <-set- false;
    return is_match;
```

and number of followers grow over time. Section 6 demonstrates that the proved privacy risk can still be exploited even when the task is more challenging.

5.2 Algorithm

In Algorithm 1 we formulate a general de-anonymization algorithm **DeHIN** to prey upon the risk of a heterogeneous information network as identified in Section 4.

The attribute values of the target entity and the entity from the auxiliary dataset is compared by function *entity_attribute_match*. This function can be configured by users depending on different scenarios. We consider the auxiliary dataset grows from the target dataset in the threat model. So some attribute values may grow over time, such as number of tweets.

The recursive Algorithm 2 is to assist DeHIN to compare the distance-n neighbors from a target entity and an entity in the auxiliary dataset whose attributes are matched with those of the target. Likewise, function *link_attribute_match* compares the attribute values of target meta paths (links in the target network schema), if any, and is configurable. The challenge lies in how to compare the neighbors of two entities, after their own entity and link attribute values are matched. Consider the case depicted in Figure 6, the target entity v'_8 is matched with entity v_9 in the auxiliary dataset for function *entity_attribute_match*, and the target's neighbor v'_5 is matched with v_1 and v_2 (entity v_9's neighbors) via the same type of link for the same function, v'_6 matched with v_2, v'_7 matched with v_3 and v_4. For a growing network, v_9 in the auxiliary dataset may be the "grown" target: v_9 itself matches v'_8 in profile attributes, v_9's neighbors v_1 and v_2 in fact are the non-anonymized v'_5 and v'_6, who are the neighbors of the target via the same type of link. Although v'_7 may be either v_3 or v_4 since they are matched via the same type of link, we can consider the remaining neighbor of v_9, either v_4 or v_3, to be the newly developed relationships during the time gap of the target and auxiliary datasets. Therefore, it is a maximum bipartite matching problem in graph theory (the candidate set for v'_5, $C(v'_5) = \{v_1, v_2\}$, $C(v'_6) = \{v_2\}$, $C(v'_7) = \{v_3, v_4\}$), and the most efficient Hopcroft-Karp algorithm is employed to decide whether such a maximum bipartite matching exists [6]. As long as a maximum bipartite matching exists (*e.g.*, v'_5, v'_6 and v'_7 match v_1, v_2 and v_3 respectively; or v'_5, v'_6 and v'_7 match v_1, v_2 and v_4 respectively), v_9 is considered as a candidate of v'_8. Finally DeHIN returns a *candidate set* containing all entities from the auxiliary dataset that may be the target entity. If the size of the correct candidate set is 1, a unique matching is found and the target entity is successfully de-anonymized.

It should be pointed out that, DeHIN is suitable for the general information network and is also applicable to a homogeneous information network, when it is considered as a special case of the general information network whose number of entity type and link type are 1. Besides, DeHIN does not employ isomorphism testing algorithms due to its high computational cost although we believe it can further enhance the accuracy. In the next section, we show

Table 1: Privacy Risk of the Anonymized t.qq Dataset (density: 0.01, size: 1000) increases as the amount of utilized target network schema link types increases (in percentage)

Types of Links \ Max. Distance	1	2	3
f	84.4	93.8	93.8
m	85.4	93.6	93.8
c	87.6	93.6	93.9
r	90.2	94.2	94.3
f-m	96.0	98.5	98.6
f-c	95.6	98.5	98.5
f-r	96.8	98.5	98.5
m-c	89.9	94.0	94.2
m-r	91.2	94.4	94.5
c-r	91.8	94.4	94.5
f-m-c	96.5	98.5	98.6
f-m-r	96.9	98.6	98.6
f-c-r	96.8	98.6	98.6
m-c-r	92.3	94.5	94.6
f-m-c-r	96.9	98.6	98.6

*f: follow; m: mention; r: retweet; c: comment
**Max. Distance n*: max. distance of utilized neighbors to target entities
*$n = 0$: only target entities' profiles are utilized and risk is always 1.1%

DeHIN is effective in the settings of a heterogeneous information network even without incorporating isomorphism tests.

6. EVALUATION

In this section, we evaluate the privacy risk and DeHIN performance on *t.qq* dataset. Then we show DeHIN is able to beat the investigated graph anonymization algorithms in the settings of a heterogeneous information network, while further sacrificing utility is able to defend the attack. It is also shown that DeHIN undermines the notion of "security by obscurity" for privacy preservation.

6.1 Case Study of t.qq Dataset

Following the motivating example in Section 1.1, we first evaluate the privacy risk as formalized in Section 4. Details of the anonymized KDD Cup 2012 *t.qq* dataset is depicted in Section 1.1 and Section 3. 500 target graphs of 1,000 vertices are sampled from *t.qq* dataset where vertices are randomly sampled and all the edges among them are preserved. Although a power-law out-degree distribution is assumed in the analysis (Section 4), since increasing privacy risk requires more edges to utilize different distances of neighbors from a target user, the privacy risk may vary when in reality heterogeneous information networks are of different densities:

$$density = \frac{|E|}{m\,|V|^2 + (|\mathcal{L}| - m)\,|V|\,(|V| - 1)} \tag{4}$$

In Equation 4, $|E|$ and $|V|$ are the number of edges and vertices in the network. $|\mathcal{L}|$ indicates the total number of link types in the network and m denotes the number of link types which allow nodes to self-link. The denominator of Equation 4 represents the maximum possible number of edges in the network and the value of density is always between 0 and 1.

57 of the sampled target graphs have density 0.01. The average cardinality of gender, yob, number of tweets, and number of tags for these 57 samples are 3, 87, 643, and 11 respectively. Considering the relatively small size of the target dataset, to better observe the growth of risk and variation in terms of different amounts of link types, only the number of tags is used in computing the entity cardinality $\mathbb{C}(\mathcal{E}^*)$. Results in Table 1 and Figure 7 (Figure 7 averages the privacy risk utilizing the same amount of link types) show that privacy risk calculated by Theorem 1 increases as the utilized heterogeneity information grows, which is the amount of target network schema link types. The drastic growth from distance 0 to 1 is consistent with the established order of growth in Equation 2 and Equation 3, then risk grows asymptotically towards 1 until it remains unchanged. Recall Section 4.5, the results also justify the practical efforts of reducing accessible link types is able to reduce $\mathbb{C}(\mathcal{L}^*)$ and hence privacy risk. When no link information is accessible, $n = 0$ and privacy risk is reduced efficiently given that the entity cardinality is not large as compared with the entity size.

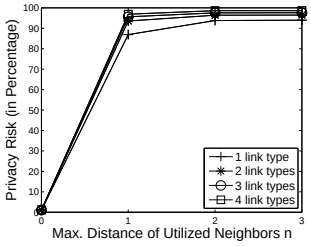


Figure 7: Privacy risk increases with more link types

To evaluate the performance of DeHIN proposed in Section 5 on *t.qq* dataset, the entire anonymized *t.qq* dataset is used as the auxiliary dataset while the target dataset is the sampled 500 target graphs and none of them contains cliques of size over 3. We will show DeHIN works effectively without the need to create any "sybil nodes" or to rely on easily-detectable graph structures in a large-scale network as required in the existing attacks [2, 12]. The anonymized user IDs (randomly assigned strings) in both target and auxiliary datasets are not used for attribute value matching. After DeHIN employs the remaining attribute and link information described in the motivating example (*user profile, mention, retweet, comment, follow data*) to establish the unique matching between the target user in the target dataset and a user in the auxiliary dataset, the anonymized user IDs will serve as the ground truth to decide if the unique matching is correct.

Since a social network generally grows over time, we intentionally consider attributes such as *tweet count*, *mention strength*, *retweet strength*, *comment strength* may grow between the time gap of the auxiliary and target datasets. Therefore, the attribute matching functions are configured to allow any user entity in the auxiliary dataset with values of these attributes greater than or equal to those of the target user to be a candidate. Likewise, we also intentionally consider links may be newly formed in the auxiliary dataset for link matching. These considerations make the de-anonymization scenario more practical and more challenging since they will potentially introduce more candidates comparing with the exact attribute or link value matching.

The entire auxiliary dataset contains 2,320,895 user entities. With random guessing, the adversary may de-anonymize a user from the target dataset with probability no higher than $\frac{1}{2{,}320{,}895}$. If the candidate size can be reduced to 100 including the target, the random guessing may be correct with a drastically increased chance of $\frac{1}{100}$. If the candidate size is exactly 1 and such a unique matching is correct, the de-anonymization is successful. Hence, we define two metrics for the experiments:

$$Precision = \frac{\sum_{i=1}^{|V'|} s(v_i')}{|V'|},$$

$$Reduction\ Rate = \frac{1}{|V'|}\sum_{i=1}^{|V'|}(1 - \frac{|C(v_i')|}{|V|}),$$

Table 2: Performance of DeHIN on t.qq anonymized dataset (in percentage)

Density	Max. Distance 0		Max. Distance 1		Max. Distance 2		Max. Distance 3	
	Precision	Reduction Rate	Precision	Reduction Rate	Precision	Reduction Rate	Precision	Reduction Rate
0.001	4.1	99.836	12.6	99.848	12.6	99.848	12.6	99.848
0.002	5.1	99.925	22	99.947	22.7	99.948	22.7	99.948
0.003	6.5	99.917	32.8	99.944	33.5	99.945	33.5	99.945
0.004	4.3	99.907	39.4	99.941	40.8	99.942	40.9	99.942
0.005	4.3	99.927	48.7	99.969	49.8	99.969	49.9	99.969
0.006	7	99.920	59.4	99.979	61.6	99.980	61.7	99.980
0.007	5.1	99.908	65.6	99.977	68.8	99.978	68.9	99.978
0.008	5.3	99.921	76.6	99.989	78.8	99.989	79	99.989
0.009	6.4	99.914	86.2	99.997	88.6	99.997	88.8	99.997
0.01	5.4	99.892	92.5	99.989	95.6	99.990	95.7	99.990

**Max. Distance n*: max. distance of utilized neighbors to target entities; when $n = 0$, only target entities' profile attributes are utilized

Table 3: Performance of DeHIN on t.qq anonymized dataset (density: 0.01) improves as the amount of utilized target network schema link types increases (in percentage)

Types of Links	Max. Distance 1		Max. Distance 2		Max. Distance 3	
	Precision	Reduction Rate	Precision	Reduction Rate	Precision	Reduction Rate
f	68.1	99.982	77.6	99.983	77.7	99.983
m	80.9	99.976	87.8	99.976	88	99.976
c	82.8	99.975	88.7	99.976	88.8	99.976
r	81.1	99.976	88.7	99.976	88.9	99.976
f-m	89.3	99.989	94.2	99.990	94.2	99.990
f-c	90.1	99.989	94.6	99.990	94.6	99.990
f-r	89.2	99.989	94.9	99.990	95	99.990
m-c	84.7	99.976	89.6	99.976	89.7	99.976
m-r	83.2	99.976	89.5	99.977	89.7	99.977
c-r	85.2	99.976	90.3	99.976	90.5	99.976
f-m-c	91.6	99.989	94.8	99.990	94.8	99.990
f-m-r	90.6	99.989	95.1	99.990	95.2	99.990
f-c-r	91.5	99.989	95.4	99.990	95.5	99.990
m-c-r	86.5	99.977	91	99.977	91.2	99.977
f-m-c-r	92.5	99.989	95.6	99.990	95.7	99.990

*f: follow; m: mention; r: retweet; c: comment
**Max. Distance n*: max. distance of utilized neighbors to target entities; when $n = 0$, only target entities' profile attributes are utilized
*$n = 0$: only target entities' profiles are utilized—precision and reduction rate are always 5.4% and 99.892%

Table 4: Performance of DeHIN on t.qq dataset of complete graph anonymity (in percentage)

Density	Max. Distance 0		Max. Distance 1		Max. Distance 2		Max. Distance 3	
	Precision	Reduction Rate	Precision	Reduction Rate	Precision	Reduction Rate	Precision	Reduction Rate
0.001	4.1	99.836	11.5	99.847	11.9	99.847	11.9	99.847
0.002	5.1	99.925	19.7	99.941	20.9	99.941	20.9	99.941
0.003	6.5	99.917	29.8	99.938	31.6	99.938	31.6	99.938
0.004	4.3	99.907	35.8	99.936	38.3	99.936	38.4	99.936
0.005	4.3	99.927	44.1	99.963	47.1	99.963	47.1	99.963
0.006	7	99.921	54.3	99.973	57.8	99.973	57.9	99.973
0.007	5.1	99.908	59.5	99.971	64.2	99.971	64.2	99.971
0.008	5.3	99.921	70.3	99.978	74.8	99.978	74.8	99.978
0.009	6.4	99.914	78.1	99.985	83.4	99.986	83.5	99.986
0.01	5.4	99.892	84.4	99.976	89.8	99.976	89.8	99.976

**Max. Distance n*: max. distance of utilized neighbors to target entities; when $n = 0$, only target entities' profile attributes are utilized

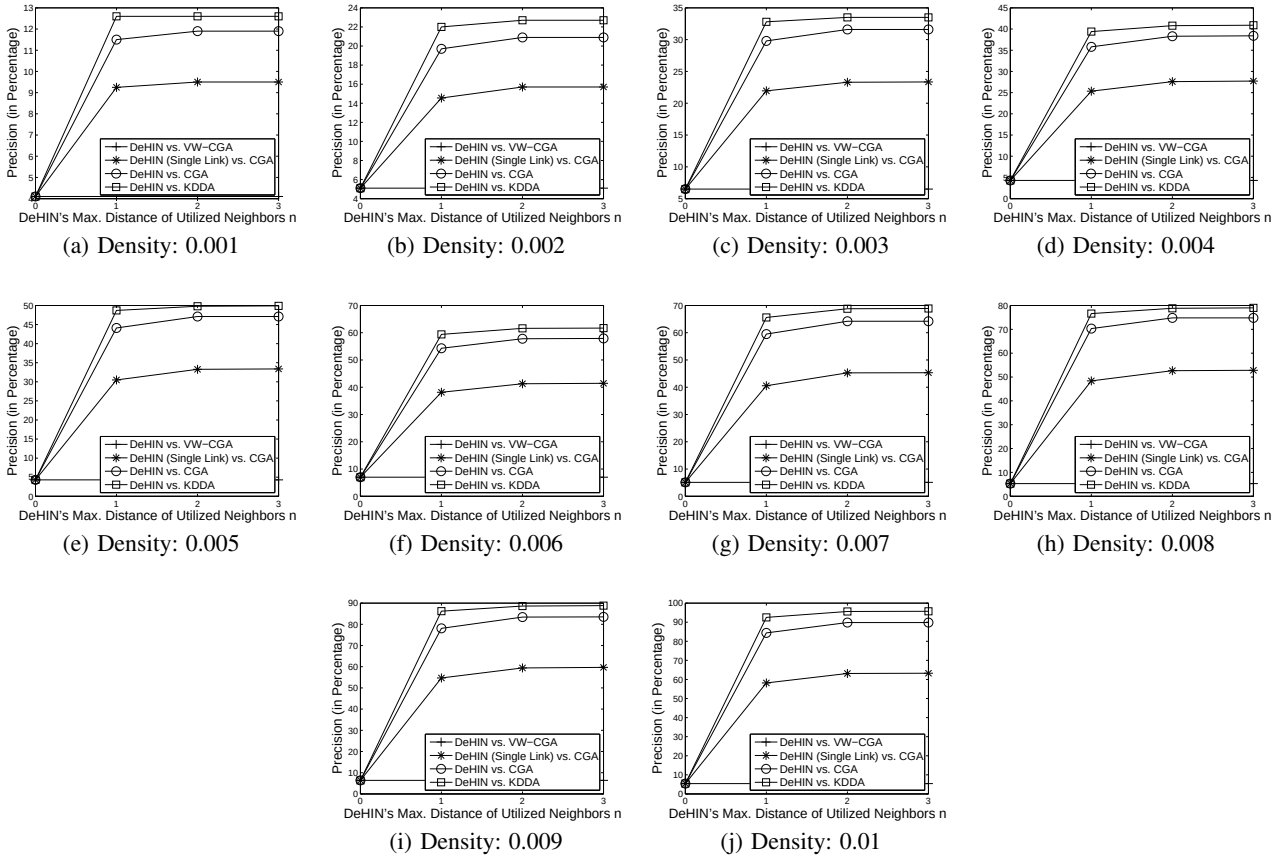
where $|V'|$ and $|V|$ are the size of the target and auxiliary dataset, $s = 1$ if $v'_i \in V'$ is successfully de-anonymized, otherwise $s = 0$, and $|C(v'_i)|$ is the size of candidate set for the target v'_i.

The performance of DeHIN on target datasets of different densities is shown in Table 2. Clearly, the general performance improves as the density of the target dataset increases because higher density indicates DeHIN may be able to utilize more neighbors to expand the dimensions of each target user to achieve unique matchings. It reveals an important problem that, if a group of people have rich social connections, they may have higher social values and may cause adversaries' attention; however, their privacy can be compromised more easily. Generally, the reduction rate looks promising as compared with the original candidate size of 2.3 million; so even when precision is relatively low on a low-density network, high reduction rate makes manual investigation of matched candidates possibly practical. For a certain density level, precision increases drastically when distance-1 neighbors are utilized, particularly for a higher-density network where there may be more neighbors. Due to the bottleneck scenarios discussed in Section 4.3 and Figure 5, the performance improves much more slowly or remains unchanged when DeHIN utilizes neighbors of longer distances.

To evaluate whether the heterogeneity of an information network improves the performance, we selectively employ different types of links in DeHIN and gradually increase the number of links in de-anonymizing the target dataset with potentially a higher social value (density = 0.01). The results in Table 3 and Figure 9 (Figure 9 averages the precision of DeHIN utilizing the same amount of link types) justifies that the performance improves as the utilized heterogeneity information grows, which is the amount of target network schema link types. Moreover, the observed growth trend is consistent to that of privacy risk in Figure 7.

6.2 Beating Complete Graph Anonymity

The utility of *t.qq* dataset has to be preserved to a certain level to



(a) Density: 0.001 (b) Density: 0.002 (c) Density: 0.003 (d) Density: 0.004

(e) Density: 0.005 (f) Density: 0.006 (g) Density: 0.007 (h) Density: 0.008

(i) Density: 0.009 (j) Density: 0.01

Figure 8: Precision of DeHIN against different anonymized heterogeneous information networks of different densities (CGA: Complete Graph Anonymity; VW-CGA: Varying Weight Complete Graph Anonymity; KDDA: KDD Cup 2012 t.qq Original Anonymization)

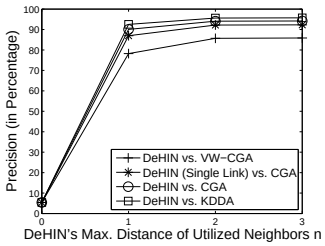


Figure 9: DeHIN Precision Improves with More Link Types

ensure effective recommendation algorithms can be designed. We now lower their utility and apply the state-of-the-art graph anonymization algorithms in Section 2.3 on *t.qq* dataset. Since adding edges to link all the users will make the entire network safer from all the structural attacks as identified in the work of k-degree, k-neighborhood, k-symmetry, k-automorphism, and k-security, to ensure the best case of defence, we formulate *complete graphs* under different types of links. *Complete Graph Anonymity* can be considered as one of the best case for the investigated graph anonymization algorithms. For instance, when the graph becomes a complete graph after fake links are added, the k turns to be the largest possible value, which is the number of vertices in the graph, for anonymization like k-degree, k-neighborhood, *etc,* as surveyed in Section 2.3. To be consistent with these original algorithms that do not consider short-circuited features and to preserve certain utility, we set short-circuited attribute values to be the same random number and keep the existing short-circuited attribute values.

To address the enhanced anonymity, DeHIN is now re-configured to remove all the links with the majority short-circuited attribute value in the entire network before taking effect. Since a social network is generally of density lower than 0.5, it can almost be ensured that all the newly added fake links will be removed from the target dataset. However, this step will mistakenly remove the real links that have the same short-circuited attribute values as the fake links from the target dataset and $\mathbb{C}(\mathcal{L}^*)$ decreases in Equation 2 and Equation 3; thus the performance of DeHIN degrades slightly as shown in Table 4 and Figure 8(a)—Figure 8(j). In Figure 8(a)—Figure 8(j), complete graph anonymity is able to lower the attack precision effectively when DeHIN only utilizes a single homogeneous link. However, DeHIN still poses great threats to complete graph anonymity, when heterogeneous links are fully utilized.

6.3 Defending DeHIN by Sacrificing Utility

To enhance preserved privacy against DeHIN, we have to further lower the utility of the target dataset by assigning randomly generated varying weights to the short-circuited attributes of each newly

added fake links. It can be observed from Figure 8(a)—Figure 8(j) that this *Varying Weight Complete Graph Anonymity* renders DeHIN ineffective when utilizing neighbors because most faked links are still preserved in the target dataset and n is clear to 0 in Equation 2 and Equation 3. However, varying weight values in the fake links cause much higher information loss than assigning the same values; thus the anonymized data utility is sacrificed much more.

6.4 "Security by Obscurity"?

While DeHIN can be launched successfully against certain anonymization (*e.g.*, DeHIN v.s. KDD Cup Original anonymization), it may be (slightly) less effective against other anonymizations (*e.g.*, complete graph anonymity) even when it is re-configured as in Section 6.2. Researchers might be tempted to suggest that, because the adversary might not know what anonymity is employed, he might not be able to launch an attack. Here, we hope to dispel this notion. Suppose an adversary always uses the re-configured DeHIN in Section 6.2, the performance on the original *t.qq* anonymization will be exactly the same as that of complete graph anonymity because likewise only the real edges of the same majority attribute values will be affected during de-anonymization. Since DeHIN still poses great threats, this is an extremely important indication that privacy preservation requires more attention from researchers.

7. CONCLUSIONS AND FUTURE WORK

Heterogeneous information networks abound in real life but privacy preservation in such new settings has not received the due attention. In this work, we defined and identified privacy risk in anonymized heterogeneous information networks and presented a new de-anonymization attack that preys upon their risk. We further experimentally substantiated the presence of privacy risk and successfully tested the attack in the KDD Cup 2012 *t.qq* dataset. One might find surprising the ease with which the devised attack can beat the investigated anonymization algorithms. While we have selected a small number of anonymization for this initial study, we have no reason to believe that other anonymization will prove impervious to this attack. Hence, our results make a compelling argument that privacy must be a central goal for sensitive heterogeneous information network publishers.

This paper presents early results of our investigation. Planned future work includes: a) explore properties of the privacy risk metric and extend its applications; b) identify possible solutions for defending DeHIN, particularly without much utility loss.

Acknowledgement

This work was supported by HHS 90TR0003-01 (SHARPS), NSF CNS 0964392 (NSF EBAM), 1017782, 1117106, 1223477, 1223495, IIS 1018723, the Adv. Digital Sci. Center UIUC, the Multimodal Info. Access and Synthesis Center UIUC, the U.S. Army Research Lab under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA) and the U.S. Army Research Office under Cooperative Agreement No. W911NF-13-1-0193. The views expressed are those of the authors only. We thank organizers of KDD Cup 2012 and Tencent Inc. for the datasets, and thank Yizhou Sun, Manish Gupta, Rui Li and Vincent Bindschaedler for insightful discussions.

8. REFERENCES

[1] http://www.kddcup2012.org/c/kddcup2012-track1.

[2] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *Proc. of the 16th intl. conference on World Wide Web*, pages 181–190, 2007.

[3] J. Cheng, A. Fu, and J. Liu. K-isomorphism: privacy preserving network publication against structural attacks. In *Proc. of intl. conf. on Management of data*, 2010.

[4] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 27–37. ACM, 2010.

[5] J. Han, Y. Sun, X. Yan, and P. Yu. Mining knowledge from data: An information network analysis approach. In *Data Engineering, 2012 IEEE 28th International Conference on*, pages 1214–1217. IEEE, 2012.

[6] J. Hopcroft and R. Karp. An n^5/2 algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 2(4):225–231, 1973.

[7] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.

[8] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008.

[9] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1):3, 2007.

[10] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, E. Shin, and D. Song. On the feasibility of internet-scale author identification. In *Security and Privacy, 2012 IEEE Symposium on*, pages 300–314. IEEE, 2012.

[11] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[12] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on*, pages 173–187. IEEE, 2009.

[13] M. Newman. *Networks: an introduction*. 2009.

[14] Y. Sun and J. Han. Mining heterogeneous information networks: Principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2), 2012.

[15] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *Intl. J. of Uncertainty, Fuz. and Knowledge-Based Sys.*, 10(05):571–588, 2002.

[16] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[17] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. 1994.

[18] W. Wu, Y. Xiao, W. Wang, Z. He, and Z. Wang. k-symmetry model for identity anonymization in social networks. In *Proceedings of the 13th international conference on extending database technology*, pages 111–122. ACM, 2010.

[19] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In *Data Engineering, 2008. IEEE 24th International Conference on*, 2008.

[20] B. Zhou, J. Pei, and W. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explorations Newsletter*, 10(2):12–22, 2008.

[21] L. Zou, L. Chen, and M. Özsu. K-automorphism: A general framework for privacy preserving network publication. *Proceedings of the VLDB Endowment*, 2(1):946–957, 2009.