Information Systems 82 (2019) 53–70

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

Privacy preserving serial publication of transactional data

Michael Bewong *, Jixue Liu, Lin Liu, Jiuyong Li

School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes Blvd, Mawson Lakes, SA 5095, Australia

HIGHLIGHTS

- A probabilistic model to calculate the risk of a serial publication is developed.
- A novel publication mechanism Sanony that prudently uses counterfeits to prevent composition attacks is developed.
- An empirical evaluation demonstrates the effectiveness of Sanony in preserving strong privacy guarantees without entirely diminishing the utility of the published data.

ARTICLE INFO

Article history:
Received 10 April 2018
Received in revised form 14 November 2018
Accepted 3 January 2019
Available online 25 January 2019
Recommended by D. Suciu

Keywords:
Privacy preservation
Serial publication
Data anonymisation
Transactional data

ABSTRACT

The continuous release of data, also called serial publication is critical for data analytics but it can lead to severe privacy disclosures *via* composition attacks. The serial publication often consists of several corpora and each corpus is an update of the previous one. While each individually published corpus may be privacy preserving, when considered together the whole serial publication may be at risk of privacy disclosures. Existing solutions addressing this problem often afford the privacy guarantees of *k*-anonymity and *l*-diversity which are prone to attribute disclosures via *skewness attacks*, and they focus only on relational data. This paper addresses the serial publication problem in the transactional data setting. First, we model the privacy disclosure risks associated with serially published data probabilistically. We then develop a rigorous privacy guarantee and a serial publication method *Sanony* that satisfies the privacy guarantee without excessive utility loss. We evaluate our method on two benchmark datasets and the results show our framework affords stronger privacy with much lower perturbation rates than existing state-of-the-art techniques.

1. Introduction

Serially published data is important for data analytics, however it can lead to privacy disclosure risks [1,2]. In such publications, the dataset is updated periodically, old transactions may be removed, and new ones may be added. Hence, each corpus in a serial publication contains only transactions in the period under consideration. For example, prisoner health data may be released every year for prisoner health research [3,4]. Table 1 illustrates such a scenario. In the table, 1b and c are successive updates of 1a involving deletions and insertions. Each table contains inmates' convictions and pre-existing health conditions. While an inmate's conviction may be common knowledge *e.g.* to a neighbour or an employer, his/her health conditions (in *italics*), referred to as private terms, must remain guarded.

The serial publication may not be privacy preserving as a whole, although each independently published corpus within it may be privacy preserving. This is illustrated by the following example.

Example 1. Suppose transactions T_1, T_2 and T_3 (Table 1) form a cluster of similar transactions in each release. Table 2a and b are *l*-diverse snapshots of Table 1a and b respectively by *anatomy* [5]. In the tables, the private and non-private terms are placed into separate partitions linked through the group IDs (G.ID). The transactions' IDs are for referencing only and not included in the publication. The chance for any individual to be linked to a private term is no more than $1/l$ ($l = 2$) making each release privacy preserving.[1] However, when the adversary knows *Laura* (T_1) was released before year 2, so her record T_1 is not in $\mathbb{T}_2$ (Table 2b), her disease can trivially be identified as *HIV* and the 2-diversity guarantee is broken. This is a form of attribute disclosure called the composition attack [2].

* Corresponding author.
E-mail addresses: michael.bewong@unisa.edu.au (M. Bewong), jixue.liu@unisa.edu.au (J. Liu), lin.liu@unisa.edu.au (L. Liu), jiuyong.li@unisa.edu.au (J. Li).

https://doi.org/10.1016/j.is.2019.01.001
0306-4379/

[1] For simplicity, we use probabilistic *l* diversity but the conclusions drawn holds for other instantiations of *l*-diversity [6].

Table 1
Serially generated prisoner health records.

(a) Original Corpus $\mathbb{T}_1$ (Year 1)

Name	Records
Laura	T_1\{theft, arson, fraud, *HIV*\}
Lucy	T_2\{theft, arson *cancer*\}
Martin	T_3\{vandal., arson\}
Shane	T_4\{abuse, arson, *cancer*\}
John	T_5\{DUI, assault, *HIV, herpes*\}
Stacy	T_6\{DUI, assault\}

(b) Original Corpus $\mathbb{T}_2$ (Year 2)

Name	Records
~~Laura~~	~~T_1\{theft, arson, fraud, *HIV*\}~~
Lucy	T_2\{theft, arson, *cancer*\}
Martin	T_3\{vandal., arson\}
Shane	T_4\{abuse, arson, *cancer*\}
~~John~~	~~T_5\{DUI, assault, *HIV, herpes*\}~~
~~Stacy~~	~~T_6\{DUI, assault\}~~
Ben*	T_7\{theft, arson, murder \}
Ivy*	T_8\{abuse, arson, mansl. \}

(c) Original Corpus $\mathbb{T}_3$ (Year 3)

Name	Records
~~Laura~~	~~T_1\{theft, arson, fraud, *HIV*\}~~
Lucy	T_2\{theft, arson, *cancer*\}
~~Martin~~	~~T_3\{vandal., arson\}~~
~~Shane~~	~~T_4\{abuse, arson, *cancer*\}~~
~~John~~	~~T_5\{DUI, assault, *HIV, herpes*\}~~
~~Stacy~~	~~T_6\{DUI, assault\}~~
Ben	T_7\{theft, arson, murder \}
Ivy	T_8\{abuse, arson, mansl. \}
Pam*	T_9\{abuse, arson, mansl. *cancer*\}
Jane*	T_{10}\{theft, arson, *HIV*\}

Table 2
2-diverse snapshots of prisoner health records.

(a) Anatomised Release of $\mathbb{T}_1$

ID	Non-private set	G.ID
T_1	{theft, arson, fraud}	1
T_2	{theft, arson}	1
T_3	{vandal., arson}	1
⋮	⋮	⋮

G.ID	disease
1	*HIV*
1	*cancer*
⋮	⋮

(b) Anatomised Release of $\mathbb{T}_2$

ID	Non-private set	G.ID
T_2	{theft, arson}	1
T_3	{vandal., arson}	1
⋮	⋮	⋮

G.ID	disease
1	*cancer*
⋮	⋮

Existing privacy preserving serial publication techniques such as [1,7–11] focus only on relational data, and afford traditional privacy guarantees of *k*-anonymity and *l*-diversity [12,13]. However it is known that these privacy guarantees are prone to attribute disclosures via *skewness attacks* [6,14]. We shall use the seminal concept of *m-invariance* in [7] to illustrate the skewness attack by attempting to handle our transactional data scenario in Example 2.

Example 2. In principle, *m-invariance* ensures that each transaction in a cluster is associated with the same set of private terms (also called *signature*) in each release to maintain *m-diversity*. This is often achieved through generalisation and the use of counterfeits. In this example, Table 3 is the 2-invariant publication of Table 1 by adapting *anatomy* [5] and foregoing generalisation.

Table 3
2-invariant prisoner health records.

(a) Published Corpus $\mathbb{T}_1^*$ (Year 1)

ID	Non-private set	G.ID	disease
Group 1			
T_1	{theft, arson, fraud}	1	*HIV*
T_2	{theft, arson}	1	*cancer*
Group 2			
T_3	{vandal., arson}	2	NULL
T_4	{abuse, arson}	2	*cancer*
Group 3			
T_5	{DUI, assault}	3	*{HIV + herpes}*
T_6	{DUI, assault}	3	NULL

(b) Published Corpus $\mathbb{T}_2^*$ (Year 2)

ID	Non-private set	G.ID	disease
Group 1			
T_2	{theft, arson}	1	*cancer*
T_c		1	*HIV*
Group 2			
T_3	{vandal., arson}	2	NULL
T_4	{abuse, arson}	2	*cancer*
T_7	{theft, arson, murder}	-	NULL
T_8	{abuse, arson, mansl.}	-	NULL

(c) Published Corpus $\mathbb{T}_3^*$ (Year 3)

ID	Non-private set	G.ID	disease
Group 1			
T_2	{theft, arson}	1	*cancer*
T_c		1	*HIV*
Group 2			
T_7	{theft, arson, murder}	2	NULL
T_9	{abuse, arson, mansl.}	2	*cancer*
Group 3			
T_8	{abuse, arson, mansl.}	3	NULL
T_{10}	{theft, arson}	3	*HIV*

The adaptation of anatomy involves the use of counterfeits, while foregoing generalisation streamlines our focus on *attribute disclosures*. In Table 3, each transaction is faithfully associated with its *signature* until its expiration. Where multiple private terms exist, we concatenate them *e.g. {HIV + herpes}*, and where there is no private term we replace with NULL. Also T_7 and T_8 (Table 3b) are deliberately suppressed for not satisfying the *m*-eligibility criterion of *m*-invariance where at most $1/m$ ($m = 2$) of new insertions can have the same private term [7]. *m-invariance* is achieved by using counterfeits (T_c), and from Table 3, we see that composition attack on *Laura* (T_1) in Table 3a is no longer possible as was the case in Example 1. This is due to the presence of the counterfeit transaction T_c in Table 3b.

Chief among the observations that can be made from this example is that the existing publication measures of [1,7–11] do not provide sufficiently strong guarantees, particularly against skewness attacks. From Example 2, comparing the probability 1/2 of John (T_5 in Table 3a) having *herpes* to the probability of anyone in the whole table having *herpes* can lead to an inference. That is, prior to knowing which group John's record belongs, his chance for having *herpes* is 1/6 (there are 6 transactions in total). After identifying his group through his non-private terms, the probability has increased significantly to 1/2 which places him at a higher risk than the rest of the population. This constitutes the skewness attack and must also be prevented [6].

Secondly, due to the underlying *m*-eligibility constraint of *m*-diversity adopted by the existing literature [7,8], their applicability to datasets that may have many common private terms is limited. This is often the case in transactional data where many transactions may have NULL private values. It may be argued that, the case

of NULL private values does not breach the m-eligibility constraint — simply for being NULL, for this we provide a counter example.

Example 3. Suppose T_7 and T_8 are published in Table 3b. An adversary who knows that Martin (T_3) and Shane (T_4) were released only after Year 2 infers the following. First, (s)he observes that T_3 and T_4 are in group 2 of Table 3a and b only. Then (s)he observes that since T_3 and T_4 are associated with NULL and *cancer* in Table 3a, they will also be associated with same in Table 3b. Now (s)he inadvertently learns that T_7 and T_8 must be associated with the remaining NULLs in Table 3b. This knowledge can then be used in the third publication (Table 3c) to breach the privacy of *Pam* (T_9) and *Jane* (T_{10}) completely.

Example 3 also illustrates an underlying attack, **transitive composition attack** (details in Section 2). This is a less direct attack and is different from *real world exclusion* [8] which relies on a different set of publication assumptions.

Thirdly, for the case where transactions have multiple private terms, the existing methods do not apply. Individuals will often have multiple health conditions and transactional data seamlessly represents this by allowing multiple private terms per transaction. This however leads to very rare combinations of private terms whose privacy becomes increasingly harder to preserve with the existing methods without excessive loss of utility *e.g.* in Table 3a, it is impossible to prevent skewness attack on John (T_5) unless all the clusters are merged by adopting *swapping and merging* techniques in [15].

It is prudent to mention that the class of *differential privacy* [16] definitions affords strong guarantees, however it still faces unresolved challenges that creep up with our serial publishing scenario. First, the concept of neighbouring datasets is often defined to mean two datasets that differ in only one record. This implicitly assumes that the adversary knows everyone else's information but the intended victim's. While this is reasonable under the assumption that the transactions are generated independently, it has been demonstrated in [17,18] that this is not always the case and can actually lead to a degradation of the privacy guarantee. Our serial publishing scenario may consist of multiple releases and each release may differ from a previous release on multiple transactions, whose generation may not be independent. Some methods [19–22] have been proposed to partly address some of these issues, but these focus only on the interactive setting, where there is a differentially private mechanism sitting between the user and the dataset, through which the user interacts with the dataset.

Second, applying differential privacy to the non-interactive setting, where an anonymised version of the dataset is released, has always been challenging. In privacy preserving data publishing, such techniques often rely on using noised contingency tables derived from counting queries. However for transactional data, such a contingency table would require $2^{|T_\rho|} - 1$ queries ($|T_\rho|$ is the number of terms in the entire publication). Clearly the complexity of such approach is not desirable. In addition, the sparsity of transactional data leads to cases with small counts and this results in very low signal-to-noise ratio since the noise added is independent of the dataset. For example, Fig. 1 shows the application of differential privacy to count statistics of the private terms derived from the original corpus in Table 1a. The published noisy counts satisfies ϵ-differential privacy ($\epsilon = 0.1$[2]) which shows very low signal-to-noise ratio. Works including [25,26] have proposed hybrid methods combining both generalisation and perturbation, however the exact relationship between generalisation and perturbation with utility is not clear. Yet still, these methods yield significantly lower utility than other primary non-interactive oriented publication mechanisms [27]. In [28], some further discussions and experiments are given to illustrate these issues.

To address the challenges, this paper extends the stronger relative privacy metric of our previous work [14] on transactional data to serial publication.[3] The metric in [14], inspired by *t-closeness* [6], ensures that the probability of linking an individual to a private term is restricted to be no more than the probability of the private term in the whole dataset by a factor of r_{th}. This is achieved via *Anony* which partitions the transactions of the dataset into clusters first; and then partitions the transactions of each cluster $\mathbf{C}$ vertically into 3 parts (each is a multiset): a non-private segment $\mathbf{C}^{\bar{S}}$, a cluster private segment C_c^S, and a global private segment C_o^S. $\mathbf{C}^{\bar{S}}$ contains the remainder of each transaction after the private terms are removed. All copies of all private terms are then put into C_c^S. The number of copies of the private terms in C_c^S may be high so as to cause privacy risks, therefore some copies are further moved to the global private segment C_o^S, to make the cluster safe. C_o^S of all clusters are then merged into the global bag C_g, a multiset which is associated with the whole corpus. This is illustrated by Example 4.

Example 4. Table 4 is the anonymised version of Table 1 for an $r_{th} = 2$ by following *Anony* [14]. In each of Table 4a, b and c, the probability of linking someone to a private term is no more than 2 times the population rate. For instance the probability of linking anyone to *herpes* in Table 4a remains at 1/6, so there is no possibility for skewness attack as was observed in Example 2.

We make the observation that there can be no composition attacks if there are no overlapping transactions across publications *i.e.* each corpus is independent of another. Consequently, existing methods for single publication will suffice. However, the serial publication problem is not trivial as overlapping transactions often do exist in reality [1,7–11]. Our proposed method, *Sanony* utilises this observation. It compares each newly updated corpus (to be published) with the previous publications to determine overlapping transactions; calculates the posterior probability of each transaction based on the overlapping transactions; and influences the overlapping transactions to feign their absence and prevent meaningful inferences. This effectively reduces the risk of transactions and prevents composition attacks.

The cost for our method, *Sanony*, to have the right privacy guarantee is a small loss of utility compared with *Anony* [14]. Our experiments in Section 4 show that in the worst case, an average increase of less than 10% in the utility loss is recorded. Specifically our contributions are as follows.

1. A probabilistic model to calculate the risk of a serial publication. This is developed by considering the adversary's posterior knowledge upon observing other publications in the serial publication. Particularly, we make use of combinatorial techniques that capture the adversary's maximal knowledge. This enables extension of stronger privacy guarantees that prevent probabilistic inferences in the single independent publication of transactional data [14] to the serial publishing scenario.
2. A novel publication mechanism *Sanony* that prudently uses counterfeits to prevent composition attacks is proposed. Our mechanism prevents composition attacks by using counterfeits to render the overlapping transactions ineffective for any inferences. We show theoretically that our publication mechanism is sound and further demonstrate that it is impermeable to minimality attacks.

[2] Typical values for ϵ range from 0.1 (weaker) to 0.01 (stronger) [23,24].

[3] Such a relative metric has been demonstrated to afford *differential privacy*-like protection under reasonable assumptions [28,29].

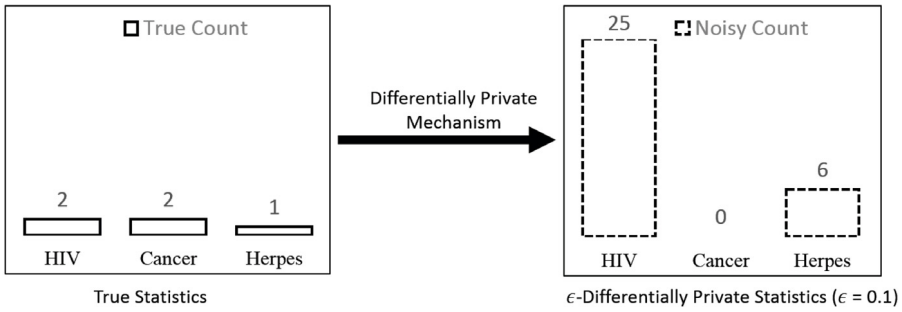


Fig. 1. Privacy preservation via differential privacy.

Table 4
Anonymised versions of prisoner health records.

(a) Anonymised Corpus $\mathbb{T}_1^{\vartheta}$ (Year 1)

$\mathbf{C}^{\bar{S}}$	C_c^S	C_g
Cluster 1 ($\mathbf{C}_{1_1}^{\vartheta}$)		
$\bar{S}_{T_1}${theft, arson, fraud}	*HIV, cancer*	
$\bar{S}_{T_2}${theft, arson}		
Cluster 2 ($\mathbf{C}_{2_1}^{\vartheta}$)		
$\bar{S}_{T_3}${vandal., arson}	*cancer*	*herpes*
$\bar{S}_{T_4}${abuse, arson}		
Cluster 3 ($\mathbf{C}_{3_1}^{\vartheta}$)		
$\bar{S}_{T_5}${DUI, assault, }	*HIV*	
$\bar{S}_{T_6}${DUI, assault}		

(b) Anonymised Corpus $\mathbb{T}_2^{\vartheta}$ (Year 2)

$\mathbf{C}^{\bar{S}}$	C_c^S	C_g
Cluster 1 ($\mathbf{C}_{1_2}^{\vartheta}$)		
$\bar{S}_{T_2}${theft, arson}	*cancer*	
$\bar{S}_{T_7}${theft, arson, murder}		
Cluster 2 ($\mathbf{C}_{2_2}^{\vartheta}$)		
$\bar{S}_{T_3}${vandal., arson}	*cancer*	
$\bar{S}_{T_4}${abuse, arson}		
$\bar{S}_{T_8}${abuse, arson, mansl.}		

(c) Anonymised Corpus $\mathbb{T}_3^{\vartheta}$ (Year 3)

$\mathbf{C}^{\bar{S}}$	C_c^S	C_g
Cluster 1 ($\mathbf{C}_{1_3}^{\vartheta}$)		
$\bar{S}_{T_2}${theft, arson}	*cancer, HIV*	
$\bar{S}_{T_7}${theft, arson, murder}		
$\bar{S}_{T_{10}}${theft, arson}		
Cluster 2 ($\mathbf{C}_{2_3}^{\vartheta}$)		
$\bar{S}_{T_8}${abuse, arson, mansl.}	*cancer*	
$\bar{S}_{T_9}${abuse, arson, mansl.}		

3. An empirical evaluation of *Sanony* on real datasets with state-of-the-art anonymisation techniques. The results demonstrates the effectiveness of *Sanony* in preserving strong privacy guarantees without entirely diminishing the utility of the published data.

To the best of our knowledge, this is the first work to extend a such a relative privacy metric to the serial publication scenario.

The rest of this paper is organised as follows: In Section 2, we develop, step-by-step, our probabilistic privacy model and then formulate the problem definition. In Section 3, we present the theory behind our solution and conclude the section with an algorithm and its analysis. Section 4 presents a rigorous empirical evaluation on two benchmark datasets and in comparison with other state-of-the-art methods. In Section 5, we review the relevant literature and describe some of the pitfalls in existing work. Finally, we present our conclusions in Section 6.

2. Preliminaries & privacy model

In this section, we develop a privacy guarantee for the serial publication of transactional data. To develop this guarantee, we first consider the relative privacy guarantees established in [14] for the single independent publication scenario and recap the publication mechanism for satisfying it.

Second, we analyse the privacy risks associated with serial publication by considering the adversary's posterior knowledge. The adversary's posterior knowledge is the adversary's belief that a transaction has a sensitive term after observing other publications in the serial publication. We propose theoretically sound approaches for calculating this posterior knowledge by relying on combinatorics and making use of the Bayes' theorem. We then further analyse our approach by considering other potential attack scenarios.

Finally, we conclude the section by presenting a formal definition of the privacy preserving serial publication problem that we seek to address.

2.1. Preliminaries

Let $\mathcal{P} = [\mathbb{T}_1, \ldots, \mathbb{T}_m]$ be a serial collection of corpora (Fig. 2) such that each corpus $\mathbb{T} \in \mathcal{P}$ is a set of transactions $\{T_1, \ldots, T_n\}$ and each transaction $T \in \mathbb{T}$ is a set of terms, *e.g.* the transaction $T_1 = \{theft, arson, fraud, HIV\}$ in corpus $\mathbb{T}_1$ of Table 1. A subset $\mathbf{C}$ of transactions in $\mathbb{T}$ is called a **cluster**. A partition $\mathbb{C}$ of the corpus $\mathbb{T}$ is a set of clusters $\mathbb{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ such that $\bigcup_{i=1}^{k} \mathbf{C}_i = \mathbb{T}$ and $\mathbf{C}_i \cap \mathbf{C}_j = \emptyset\ (i \neq j)$. $T_{\mathcal{P}}$ is the set of all terms in the collection $\mathcal{P}$. $S_{\mathcal{P}}$ is the set of all private terms in $T_{\mathcal{P}}$ and $\bar{S}_{\mathcal{P}} = T_{\mathcal{P}} \setminus S_{\mathcal{P}}$ is the set of all non-private terms in $\mathcal{P}$. With $S_{\mathcal{P}}$, a transaction is often divided into two sets: the **non-private (term) set** $\bar{S}_T$ containing the non-private terms and the **private (term) set** S_T containing the private terms. If a transaction contains $s \in S_{\mathcal{P}}$ it is called an **s-transaction**, else it is an **$\bar{s}$-transaction** (**non-s**). The function $N(s, \mathcal{X})$ returns the number of copies of the s term in the collection $\mathcal{X}$ while $N(\mathcal{X})$ returns the number of transactions in $\mathcal{X}$. For example, $N(s, \mathbf{C})$ is the number of s-transactions while $N(\mathbf{C})$ is the cardinality $|\mathbf{C}|$.

Now we wish to publish a new corpus $\mathbb{T}_{m+1}$ to update $\mathcal{P}^* = [\mathbb{T}_1^*, \ldots, \mathbb{T}_m^*]$ which is already published (Fig. 3). We assume the publication process involves 2 steps: (1) $\mathbb{T}_{m+1}$ (raw dataset) is anonymised to $\mathbb{T}_{m+1}^{\vartheta}$ via *Anony* [14] and $\mathbb{T}_{m+1}^{\vartheta}$ is privacy preserving by itself; (2) we apply *Sanony*, our method to be presented in Section 3, to transform $\mathbb{T}_{m+1}^{\vartheta}$ to $\mathbb{T}_{m+1}^*$ so that $\mathcal{P}^* + [\mathbb{T}_{m+1}^*] = [\mathbb{T}_1^*, \ldots, \mathbb{T}_m^*, \mathbb{T}_{m+1}^*]$ as a whole is privacy preserving. Consequently the superscript ϑ denotes the independent publication (anonymisation) of $\mathbb{T}_{m+1}$ and $*$ denotes the final (serial) publication of $\mathbb{T}_{m+1}$. Table 5 summarises the frequently used notations.

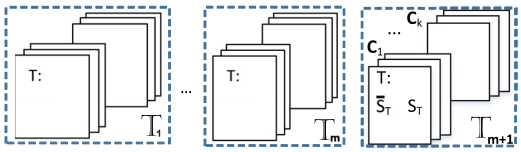


Fig. 2. Serial corpora $\mathcal{P} = [\mathbb{T}_1, \ldots, \mathbb{T}_m, \mathbb{T}_{m+1}]$.

Table 5
Summary of notations.

Notation	Meaning
T	A set of terms $\{t_1, \ldots, t_v\}$ called a transaction
$\mathbf{C}$	A subset of transactions $\{T_1, \ldots, T_{\|\mathbf{C}\|}\}$ called a cluster
$\mathbb{T}$	A set of transactions $\{T_1, \ldots, T_n\}$ called a corpus
$\mathcal{P}$	A serial corpora $[\mathbb{T}_1, \ldots, \mathbb{T}_m]$
$\mathbb{C}$	A partition $\{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ of $\mathbb{T}$
$T_{\mathcal{P}}$	A set of all terms in $\mathcal{P}$
$S_{\mathcal{P}}$	A set of all private terms in $\mathcal{P}$
$\bar{S}_{\mathcal{P}}$	A set of all non-private terms in $\mathcal{P}$
S_T	The private term set of a transaction T
$\bar{S}_T$	The non-private term set of a transaction T
$\mathcal{X}$	Original version, where $\mathcal{X} \in \{T, \mathbf{C}, \mathbb{T}, \mathbb{C}, \mathcal{P}\}$
$\mathcal{X}^{\vartheta}$	Anonymised version of $\mathcal{X}$
$\mathcal{X}^*$	Final published version of $\mathcal{X}$
$N(s, \mathcal{X})$	The support of the term s in $\mathcal{X}$
$N(\mathcal{X})$	The number of transactions in $\mathcal{X}$

Definition 1 (*Adversary*)**.** The adversary's knowledge about a victim v consists of the victim's non-private set $\bar{S}_{T_v}$ and some published corpora to which $\bar{S}_{T_v}$ belongs, as well as the method used in publishing the data.

The adversary aims to link the victim to $\bar{S}_T$ of transaction T through $\bar{S}_{T_v}$ like in [1,7–11]. The following example illustrates how the adversary uses his knowledge to form a composition attack.

Example 5. Consider Table 4, suppose the adversary knows Laura (victim) was released after year 1 of her *theft, arson* and *fraud* charges ($\bar{S}_{T_v}$). The adversary can link $\bar{S}_{T_v}$ to the non-private set $\bar{S}_{T_1}$ in $\mathbf{C}^{\vartheta}_{1_1}$ (Table 4a). $\bar{S}_{T_1}$ is not in the second release (Table 4b), but $\bar{S}_{T_2} = \{theft, arson\}$ is in both $\mathbf{C}^{\vartheta}_{1_1}$ (Table 4a) and $\mathbf{C}^{\vartheta}_{1_2}$ (Table 4b). By considering the two releases, the absence of *HIV* in $\mathbf{C}^{\vartheta}_{1_2}$ reveals that Laura with $\bar{S}_{T_1}$ must have *HIV* in $\mathbf{C}^{\vartheta}_{1_1}$.

An individual can have a single transaction per corpus but may have multiple transactions across subsequent publications *e.g.* an inmate is released and re-convicted of another charge. This does not affect our attribute disclosure problem[4] because the linkage between the private and non-private terms of a transaction is unaffected by the number of transactions a single individual has in multiple publications. At the same time, we are not specific about the case where values of a transaction are updated over time because from the adversary's point of view such updates are essentially an operation of a deletion and an insertion of a transaction. For example, supposing an individual has *HIV* during a previous data release but is discovered not to have *HIV* in the subsequent data release (due to a previous wrong diagnosis), then the individual's previous transaction T_i will be associated with the private term *HIV* in a previous release. In the subsequent data release, T_i is updated to T_j without the term *HIV*. To the adversary, the updated transaction T_j is just a newly inserted transaction whose non-private terms happen to match his victim while the previous transaction T_i has been deleted. In such instances our work also handles this to prevent the false attribution of *HIV* to the victim.

In this work a transparent publication mechanism is assumed *i.e.* the public knows the anonymisation technique used. In Section 3, we show that the adversary is incapable of using this knowledge to form any minimality attack [14,32,33] after *Sanony* is applied. In the following, we first define our guarantee on a single corpus $\mathbb{T}$ and then use the guarantee to produce a publication $\mathbb{T}^{\vartheta}$ of $\mathbb{T}$. We then formalise the risk associated with publishing $\mathbb{T}^{\vartheta}$ in the presence of its previous publications and then present the problem definition. Subsequently, we will modify $\mathbb{T}^{\vartheta}$ to produce $\mathbb{T}^*$ by considering the interaction of $\mathbb{T}^{\vartheta}$ with its previous publications.

2.2. Single independent publication

We introduce **s-preserving** guarantee, defined on *s*-**risk**, that prevents attribute disclosures [14]. It requires the relative frequency of a private term s in a cluster $\mathbf{C}$ not to exceed that of the population by more than a user defined threshold r_{th}.

Definition 2 (*s-Risk*)**.** Given a cluster $\mathbf{C} \subseteq \mathbb{T}$ and its anonymised version $\mathbf{C}^{\vartheta}$ (formally defined later), the s-Risk γ_s of $\mathbf{C}^{\vartheta}$ is the ratio of the probability of s in $\mathbf{C}^{\vartheta}$ to the probability of s in $\mathbb{T}$:

$$\gamma_s(\mathbf{C}^{\vartheta}) = \frac{N(s, \mathbf{C}^{\vartheta})/N(\mathbf{C}^{\vartheta})}{N(s, \mathbb{T})/N(\mathbb{T})} = \frac{ss(s, \mathbf{C}^{\vartheta})}{ss(s, \mathbb{T})} \quad (1)$$

$N(s, \mathbf{C}^{\vartheta})$ and $N(s, \mathbb{T})$ are the number of s-transactions in $\mathbf{C}^{\vartheta}$ and $\mathbb{T}$ respectively; $N(\mathbf{C}^{\vartheta})$ and $N(\mathbb{T})$ are the number of transactions in $\mathbf{C}^{\vartheta}$ and $\mathbb{T}$ respectively.

The probability $ss(s, \mathbf{C}^{\vartheta}) = N(s, \mathbf{C}^{\vartheta})/N(\mathbf{C}^{\vartheta})$ is also called the **cluster rate**, and $ss(s, \mathbb{T}) = N(s, \mathbb{T})/N(\mathbb{T})$ is the **population rate** as it is over all the transactions and assumed to be public knowledge *e.g.* the *HIV* rate in a community.

[4] We focus on preventing the linkage between private terms and the non-private terms (attribute disclosure) since it is of more significance [30,31]. If a user wishes to prevent the unique identification of non-private terms (identity disclosure), a k-anonymised dataset [13] can be assumed from this point on.

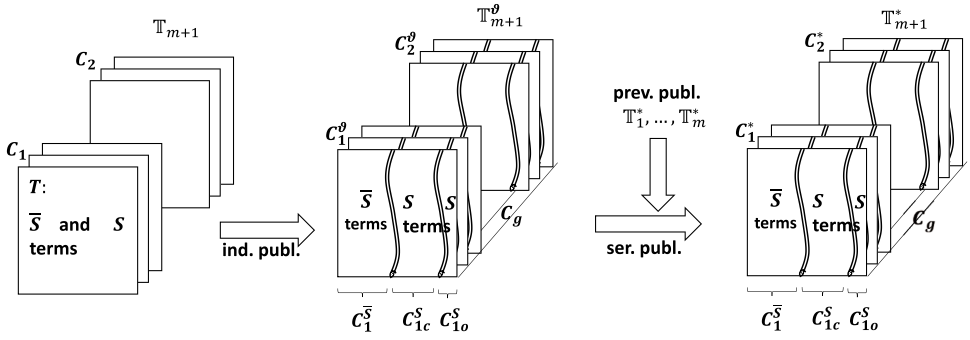


Fig. 3. Serial Publication of $\mathbb{T}_{m+1}$.

Definition 3 (*s-Preserving*)**.** Let $r_{th} \in [1, N(\mathbb{T})]$ be a user defined threshold. Given a private term s and an anonymised cluster $\mathbf{C}^{\vartheta} \subseteq \mathbb{T}^{\vartheta}$, $\mathbf{C}^{\vartheta}$ is said to be s-preserving if its s-risk γ_s is no more than r_{th}.

$$\gamma_s(\mathbf{C}^{\vartheta}) \leq r_{th} \tag{2}$$

The cluster $\mathbf{C}^{\vartheta}$ is said to be privacy preserving if it is s-preserving for every private term $s \in S_{\mathcal{P}}$.

Definition 3 above ensures that the probability for an anonymised transaction $T^{\vartheta} \in \mathbf{C}^{\vartheta}$ to have the private term s is always bound to the population rate by r_{th}. To satisfy the s-preserving criterion, we present the **anonymisation** steps of *Anony* [14], already illustrated in Example 4, as follows.

Given a user defined threshold r_{th}, a corpus $\mathbb{T}$ and its partition $\mathbb{C}$, the anonymisation of $\mathbb{C}$ has three main steps, *segregation*, *sanitisation*, and *refining*, which is applied to each cluster $\mathbf{C} \in \mathbb{C}$ to get $\mathbf{C}^{\vartheta}$.

1. *Segregation* vertically partitions $\mathbf{C}$ into two segments $\mathbf{C}^{\bar{S}}$, a multiset of the non-private sets of all transactions in $\mathbf{C}$; and $C^S_{\uplus}$, a multiset union of the private sets of all transactions in $\mathbf{C}$. This breaks the link between non-private and private terms of a transaction, making it probabilistic.
2. *Sanitisation* partitions the multiset $C^S_{\uplus}$ of private terms into two multisets, the cluster private segment C^S_c, and the global private segment C^S_o; moving any term from C^S_o to C^S_c will cause a violation of the criterion $\gamma_s(\mathbf{C}^{\vartheta}) \leq r_{th}$. The calculation of $\gamma_s(\mathbf{C}^{\vartheta})$ when $\mathbf{C}$ is partitioned into $\mathbf{C}^{\vartheta} = \{\mathbf{C}^{\bar{S}}, C^S_c, C^S_o\}$ is given by Formula (4).
3. Repeat steps 1 and 2 for every cluster $\mathbf{C} \in \mathbb{C}$. The multiset union of the global private segment C^S_o of all clusters becomes the **global bag** of private terms denoted $C_g = \biguplus_{j=1}^{k} C^S_{o_j}$ which is associated with the whole corpus.
4. *Refining* ensures that if any cluster $\mathbf{C}^{\vartheta}$ satisfies the *exclusion condition* $\gamma_s(\mathbf{C}^{\vartheta}) < r_{th}$[5] some copies of s are moved from the global bag C_g into C^S_c (of $\mathbf{C}^{\vartheta}$) until the exclusion condition is violated or no copies of s remain in C_g. It ensures that the knowledge of the publishing mechanism does not lead to a minimality attack [14,32]. In particular, the adversary may use the criterion $\gamma_s(\mathbf{C}^{\vartheta}) < r_{th}$ derived from the sanitisation step to exclude some clusters from being associated with the private terms in the global bag, and to increase the chance of other clusters having those private terms. That is, if the addition of a copy of the private term s from C_g to C^S_c does not violate the preserving criterion $\gamma_s(\mathbf{C}^{\vartheta}) \leq r_{th}$, then C^S_c cannot be the source for any copy of s moved to C_g as otherwise it will be a violation of the *minimality principle* [32]. (The interested reader is referred to [14,32] for more details)

After segregation, the number of s-transactions $N(s, \mathbf{C}^{\vartheta})$ in Definition 2 becomes the number of s terms $N(s, C^S_{\uplus})$. After sanitisation and refining, the number of s-transactions $N(s, \mathbf{C}^{\vartheta})$ becomes the number of s terms $N(s, C^S_c)$ plus the number of s terms in the global private segment $N(s, C^S_o)$. Since all global private segments are merged into the global bag C_g we use the following formula $J_s(\mathbf{C}^{\vartheta})$ to estimate $N(s, C^S_o)$ and consequently calculate $N(s, \mathbf{C}^{\vartheta})$ as follows.

$$N(s, \mathbf{C}^{\vartheta}) = N(s, C^S_c) + J_s(\mathbf{C}^{\vartheta}) \tag{3}$$

where $J_s(\mathbf{C}^{\vartheta}) = \frac{N(\mathbf{C}^{\vartheta}) \times N(s, C_g)}{N(\mathbb{T})}$ (see note[6]).

[5] The exclusion condition uses $\gamma_s(\mathbf{C}^{\vartheta}) < r_{th}$ and NOT $\gamma_s(\mathbf{C}^{\vartheta}) \leq r_{th}$ since when $\gamma_s(\mathbf{C}^{\vartheta}) = r_{th}$, it is not possible to add any s terms from the global bag without violating the privacy guarantee. Consequently $\mathbf{C}^{\vartheta}$ is not a candidate to be excluded from owning any of the s terms in the global bag

[6] In subsequent sections we shall consider the integer values of $J_s(\mathbf{C}^{\vartheta})$ as $Round(J_s(\mathbf{C}^{\vartheta}))$ to avoid representing the number of private terms as fractions.

The number of private terms $J_s(\mathbf{C}^{\vartheta})$ from C_g is calculated probabilistically from the ratio of the size of the cluster to the size of the corpus $N(\mathbf{C}^{\vartheta})/N(\mathbb{T})$ and the number of copies of s in C_g, $N(s, C_g)$. The risk $\gamma_s(\mathbf{C}^{\vartheta})$ in Definition 2 becomes:

$$\gamma_s(\mathbf{C}^{\vartheta}) = \frac{\big(N(s, C^S_c) + J_s(\mathbf{C}^{\vartheta})\big)/N(\mathbf{C}^{\vartheta})}{ss(s, \mathbb{T})} \tag{4}$$

$\mathbf{C}^{\vartheta} = \{\mathbf{C}^{\bar{S}}, C^S_c, C^S_{o'}\}$ is the anonymised cluster where $C^S_{o'}$, the multiset of s terms derived from $J_s(\mathbf{C}^{\vartheta})$ for every term $s \in S$, approximates C_o *i.e.* $N(s, C^S_{o'}) = J_s(\mathbf{C}^{\vartheta})$ $(s \in S)$. The anonymisation of each cluster $\mathbf{C} \in \mathbb{C}$ for every private term $s \in S_{\mathcal{P}}$ becomes the anonymised corpus $\mathbb{T}^{\vartheta}$. $\mathbb{T}^{\vartheta}$ is privacy preserving since the criterion $\gamma_s(\mathbf{C}^{\vartheta}) \leq r_{th}$ is satisfied for every cluster $\mathbf{C}^{\vartheta} \in \mathbb{C}^{\vartheta}$ and there can be no risk from the C_g since it is globally shared by all transactions in the corpus so Formula (2) is guaranteed. If the anonymised cluster $\mathbf{C}^{\vartheta} = \{\mathbf{C}^{\bar{S}}, C^S_c, C^S_{o'}\}$ is published without any further modifications, it becomes the published version $\mathbf{C}^*$. It is obvious that after anonymisation, $ss(s, \mathbb{T}) = ss(s, \mathbb{T}^{\vartheta})$ in Formula (4) since no private terms were added or deleted.

In $\mathbf{C}^* = \{\mathbf{C}^{\bar{S}}, C^S_c, C^S_{o'}\}$, non-private terms and the private terms of a transaction are only linked probabilistically. This probability given by Definition 4 is used later to determine the disclosure risk in a serial publication.

Definition 4 (*Random Reconstruction*)**.** Given $T \in \mathbf{C}$ and its published cluster $\mathbf{C}^* = \{\mathbf{C}^{\bar{S}}, C^S_c, C^S_{o'}\}$, the reconstruction T^r of T from $\mathbf{C}^*$ is computed by assigning s to $\bar{S}_T \in \mathbf{C}^{\bar{S}}$ with a probability $N(s, \mathbf{C}^*)/N(\mathbf{C}^*)$ where $N(s, \mathbf{C}^*) = N(s, C^S_c) + N(s, C^S_{o'})$.

2.3. Privacy disclosures in a serial publication

In a serial publication $\mathcal{P}^*$, composition attacks are possible only if some transactions of $\mathbb{T}^* \in \mathcal{P}^*$[7] are common to at least its preceding $(\mathbb{T}^*_{-1})$ or succeeding $(\mathbb{T}^*_{+1})$ publication as previously noted. The common transactions, called **overlap** (Definition 6), are also fundamental for the attack scenarios handled in [1,7–11].

The overlap between any two original clusters $\mathbf{C}$ and $\mathbf{C}'$ is their common transactions $\mathbf{C} \cap \mathbf{C}'$. When they are anonymised and published as $\mathbf{C}^* = \{\mathbf{C}^{\bar{S}}, C^S_c, C^S_{o'}\}$ and $\mathbf{C}'^* = \{\mathbf{C}'^{\bar{S}}, C'^S_c, C'^S_{o'}\}$, the common transactions $\mathbf{C}^{\bar{S}} \cap \mathbf{C}'^{\bar{S}}$ involves only the non-private sets. To increase his/her belief about the private term of any individual, the adversary must utilise the overlap by linking the private terms $s \in S_{\mathcal{P}}$ to the common non-private sets. As the links are probabilistic, (s)he can only identify some possible copies of s to link to the common non-private sets (Example 6).

Example 6. The common non-private set between $\mathbf{C}^{\vartheta}_{1_1}$ and $\mathbf{C}^{\vartheta}_{1_2}$ in Table 4 is $\{\bar{S}_{T_2}\}$. By observation, it has no *HIV* terms *i.e.* in $\mathbf{C}^{\vartheta}_{1_1}$, $\{\bar{S}_{T_2}\}$ can have 0 or 1 copy of *HIV* but it can have at most 0 copies of *HIV* in $\mathbf{C}^{\vartheta}_{1_2}$.

The possible number of copies of a private term that can be linked to an overlap is formally developed in Proposition 1 by considering the s-**range** as follows.

Definition 5 (*s-Range*)**.** Let $\mathbf{C}^* = \{\mathbf{C}^{\bar{S}}, C^S_c, C^S_{o'}\}$ and $\mathbf{C}'^* = \{\mathbf{C}'^{\bar{S}}, C'^S_c, C'^S_{o'}\}$ be any two published clusters in two different publications and $\mathbf{O}^{\bar{S}}(\mathbf{C}^*, \mathbf{C}'^*) = \mathbf{C}^{\bar{S}} \cap \mathbf{C}'^{\bar{S}}$ (shortened $\mathbf{O}^{\bar{S}}$) be their common non-private set. Given the private term $s \in S_{\mathcal{P}}$, the number of copies of s shared by $\mathbf{O}^{\bar{S}}$ w.r.t. $\mathbf{C}^*$ is in the range $[f_{1_{\mathbf{C}^*}}, f_{2_{\mathbf{C}^*}}]_s$:

$$f_{1_{\mathbf{C}^*}} = f_1(\mathbf{C}^*, \mathbf{O}^{\bar{S}}) = \max\Big\{\big(N(s, \mathbf{C}^*) - N(\mathbf{C}^{\bar{S}} \setminus \mathbf{O}^{\bar{S}})\big), 0\Big\}$$

[7] Subsequently, we use $T^* \in \mathcal{P}^*$ to mean a transaction T^* in any published corpus $\mathbb{T}^* \in \mathcal{P}^*$. Similarly, we use $\mathbf{C}^* \in \mathcal{P}^*$ to mean a published cluster $\mathbf{C}^*$ in any published corpus $\mathbb{T}^* \in \mathcal{P}^*$.

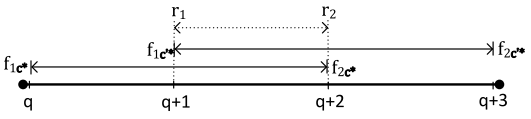


Fig. 4. *s*-range of an overlap.

$$f_{2_{\mathbf{C}^*}} = f_2(\mathbf{C}^*, \mathbf{O}^{\bar{S}}) = \min\{N(\mathbf{O}^{\bar{S}}), N(s, \mathbf{C}^*)\} \quad (5)$$

$[f_{1_{\mathbf{C}'^*}}, f_{2_{\mathbf{C}'^*}}]$ is also calculated by substituting $\mathbf{C}^*$ with $\mathbf{C}'^*$.

$f_{1_{\mathbf{C}^*}}$ is the minimum copies of s that can be linked to $\mathbf{O}^{\bar{S}}$ w.r.t. $\mathbf{C}^*$. It is developed by first linking the maximum possible copies of s to the non-overlapping transactions $\mathbf{C}^{\bar{S}} \setminus \mathbf{O}^{\bar{S}}$ of $\mathbf{C}^*$ given by $N(\mathbf{C}^{\bar{S}} \setminus \mathbf{O}^{\bar{S}})$; the remainder $N(s, \mathbf{C}^*) - N(\mathbf{C}^{\bar{S}} \setminus \mathbf{O}^{\bar{S}})$ is then linked to $\mathbf{O}^{\bar{S}}$ as the minimum copies of s that $\mathbf{O}^{\bar{S}}$ can have. The use of $\max\{(N(s, \mathbf{C}^*) - N(\mathbf{C}^{\bar{S}} \setminus \mathbf{O}^{\bar{S}})), 0\}$ ensures non-negative values only and $f_{1_{\mathbf{C}^*}} = 0$ means $\mathbf{O}^{\bar{S}}$ may have no copies of s. $f_{2_{\mathbf{C}^*}}$ is the maximum copies of s that can be linked to $\mathbf{O}^{\bar{S}}$ w.r.t. $\mathbf{C}^*$. It is derived as the maximum possible copies of s from $N(s, \mathbf{C}^*)$ that can be linked to $\mathbf{O}^{\bar{S}}$. This is given by the minimum of the number of overlapping transactions $N(\mathbf{O}^{\bar{S}})$ and the number of s terms $N(s, \mathbf{C}^*)$ *i.e.* $\min\{N(\mathbf{O}^{\bar{S}}), N(s, \mathbf{C}^*)\}$.

Proposition 1. *Let $\mathbf{C}^*$ and $\mathbf{C}'^*$ be any two published clusters and $\mathbf{O}^{\bar{S}}$ be their common non-private sets. Given a private term $s \in S_{\mathcal{P}}$ and the s-range $[f_{1_{\mathbf{C}^*}}, f_{2_{\mathbf{C}^*}}]_s$ w.r.t. $\mathbf{C}^*$ and $[f_{1_{\mathbf{C}'^*}}, f_{2_{\mathbf{C}'^*}}]_s$ w.r.t. $\mathbf{C}'^*$ (Definition 5), the number of copies of s shared by $\mathbf{O}^{\bar{S}}$ is in $[r_1, r_2]_s$, where $[r_1, r_2]_s$ is the intersection of $[f_{1_{\mathbf{C}^*}}, f_{2_{\mathbf{C}^*}}]_s$ and $[f_{1_{\mathbf{C}'^*}}, f_{2_{\mathbf{C}'^*}}]_s$:*

$$[r_1, r_2]_s = [f_{1_{\mathbf{C}^*}}, f_{2_{\mathbf{C}^*}}]_s \cap [f_{1_{\mathbf{C}'^*}}, f_{2_{\mathbf{C}'^*}}]_s \quad (6)$$

Given an ordering $\{f_{1_{\mathbf{C}^*}} \prec f_{1_{\mathbf{C}'^*}} \prec f_{2_{\mathbf{C}^*}} \prec f_{2_{\mathbf{C}'^*}}\}$ of the set $\{f_{1_{\mathbf{C}^*}}, f_{2_{\mathbf{C}^*}}, f_{1_{\mathbf{C}'^*}}, f_{2_{\mathbf{C}'^*}}\}$, r_1 and r_2 are the 2nd and 3rd values of the ordered set respectively (Fig. 4).

In Proposition 1, r_1 and r_2 are the minimum and maximum number of copies of s shared by $\mathbf{O}^{\bar{S}}$ respectively based on both $\mathbf{C}^*$ and $\mathbf{C}'^*$.

It is possible that $[r_1, r_2]_s = \emptyset$ when $[f_{1_{\mathbf{C}^*}}, f_{2_{\mathbf{C}^*}}]_s$ and $[f_{1_{\mathbf{C}'^*}}, f_{2_{\mathbf{C}'^*}}]_s$ do not intersect. This can occur when there is a substitution of some transactions between subsequent publications *e.g.* consider the published cluster $\mathbf{C}^* = \{\{\bar{S}_T = A\}, \{s_1\}, \{\}\}$ in $\mathbb{T}^*$ representing $\mathbf{C}^{\bar{S}}$, C_c^S and $C_{o\prime}^S$ respectively; and another published cluster $\mathbf{C}'^* = \{\{\bar{S}_{T'} = A\}, \{\}, \{\}\}$ in a different publication $\mathbb{T}'^*$ such that $\bar{S}_T$ in $\mathbb{T}^*$ was replaced by $S_{T'}$ in $\mathbb{T}'^*$. Obviously the overlapping non-private set is $\mathbf{O}^{\bar{S}} = \{A\}$ and by calculation, $[f_{1_{\mathbf{C}^*}} = 1, f_{2_{\mathbf{C}^*}} = 1]_{s_1}$ and $[f_{1_{\mathbf{C}'^*}} = 0, f_{2_{\mathbf{C}'^*}} = 0]_{s_1}$ (Formula (5)). The range $[r_1, r_2]_{s_1} = \emptyset$ and the adversary is not benefited by observing the overlap.

In the above example, the adversary identifies that $\bar{S}_T$ has s_1 only because s_1 is a trivial private term. Trivial because, prior to observing the second publication $\mathbb{T}'^*$, the chance of linking s_1 to $\bar{S}_T$ in $\mathbf{C}^*$ is 100%. In our s-preserving definition (Definition 3), trivial private terms occur only if the population rate $ss(s, \mathbb{T})$ and the user allowed risk threshold r_{th} are sufficiently high *i.e.* $ss(s, \mathbb{T}) \cdot r_{th} \geq 1$.

Definition 6 (*Overlap*)**.** Let $\mathbf{C}^* \in \mathbb{T}^*$ and $\mathbf{C}'^* \in \mathbb{T}'^*$ be any two published clusters of $\mathcal{P}^*$, the overlap $\mathbf{O}(\mathbf{C}^*, \mathbf{C}'^*)$ (shortened $\mathbf{O}$) of $\mathbf{C}^*$ and $\mathbf{C}'^*$ is defined to have two parts $\mathbf{O} = \{\mathbf{O}^{\bar{S}}, \mathsf{O}^S\}$:

$$\mathbf{O}^{\bar{S}} = \mathbf{C}^{\bar{S}} \cap \mathbf{C}'^{\bar{S}} \qquad \mathsf{O}^S = \{[r_1, r_2]_s \mid s \in S_{\mathcal{P}}\} \quad (7)$$

$\mathbf{O}^{\bar{S}}$ is the (non-empty) common non-private sets in $\mathbf{C}^*$ and $\mathbf{C}'^*$, O^S is a set of integer pairs, each being the range of possible copies of $s \in S_{\mathcal{P}}$ shared by $\mathbf{O}^{\bar{S}}$ given by Formula (6).

When $r_1 = r_2 = 0$, the overlap $\mathbf{O}$ does not have s even though each of the clusters $\mathbf{C}^*$ and $\mathbf{C}'^*$ may contain s. In subsequent presentation, such ranges are omitted for brevity.

Example 7. By Definition 6, the overlap of $\mathbf{C}^{\vartheta}_{1_1}$ and $\mathbf{C}^{\vartheta}_{1_2}$ (Table 4) is $\{\{\bar{S}_{T_2}\}, \{[0, 1]_{cancer}\}\}$ as its *cancer*-range is $[0, 1]_{cancer}$ each from $\mathbf{C}^{\vartheta}_{1_1}$ and $\mathbf{C}^{\vartheta}_{1_2}$ (Definition 5). Also the overlap between $\mathbf{C}^{\vartheta}_{2_1}$ and $\mathbf{C}^{\vartheta}_{2_2}$ is $\{\{\bar{S}_{T_3}, \bar{S}_{T_4}\}, \{[1, 1]_{cancer}\}\}$.

The overlap $\{\{\bar{S}_{T_2}\}, \{[0, 1]_{cancer}\}\}$ in Example 7 leads to a direct inference that Laura ($\bar{S}_{T_1}$) has *HIV* because $\bar{S}_{T_2}$ certainly has no *HIV*, but the overlap $\{\{\bar{S}_{T_3}, \bar{S}_{T_4}\}, \{[1, 1]_{cancer}\}\}$ leads to a less direct inference called **transitive composition attack** (illustrated in Example 3). For clarity, Example 3 is re-presented as follows.

Example 8. From Example 7, it is seen that $\{\bar{S}_{T_3}, \bar{S}_{T_4}\}$ has the term *cancer*, therefore $\bar{S}_{T_8}$ cannot have *cancer* in $\mathbf{C}^{\vartheta}_{2_2}$ (Table 4b). This knowledge can then be used to infer that $\bar{S}_{T_9}$ in $\mathbf{C}^{\vartheta}_{2_3}$ (Table 4c) has *cancer* with 100% probability.

Apparently, it is necessary to consider the possibility of **derived clusters** leading to composition attacks transitively.

Definition 7 (*Derived Cluster*)**.** Let $\mathbf{O}$ be the overlap of the published clusters $\mathbf{C}^* = \{\mathbf{C}^{\bar{S}}, C_c^S, C_{o\prime}^S\}$ and $\mathbf{C}'^*\{\mathbf{C}'^{\bar{S}}, C_c'^S, C_{o\prime}'^S\}$ with the range $[r_1, r_2]_s$ for a given private term $s \in S_{\mathcal{P}}$, the **derived cluster** $\mathbf{d}(\mathbf{C}^*, \mathbf{O})$ (shortened $\mathbf{d}$) w.r.t. $\mathbf{C}^*$ has two parts, $\mathbf{d} = \{\mathbf{d}^{\bar{S}}, \mathsf{d}^S\}$:

$$\mathbf{d}^{\bar{S}} = \mathbf{C}^{\bar{S}} \setminus \mathbf{O}^{\bar{S}} \qquad \mathsf{d}^S = [r_{1_f}, r_{2_f}]_s \quad (8)$$

where $r_{1_f} = \max\{(N(s, \mathbf{C}^*) - r_2), 0\}$ and $r_{2_f} = \max\{(N(s, \mathbf{C}^*) - r_1), 0\}$

$\mathbf{d}^{\bar{S}}$ is the non-overlapping transactions of $\mathbf{C}^*$. d^S is the range of possible number of copies of s that can be linked with $\mathbf{d}^{\bar{S}}$ and it is derived with the range $[r_1, r_2]_s$ of $\mathbf{O}$. The derived cluster defined above becomes a "new" cluster which can then be used to form further overlaps with other clusters for potential composition attacks. In Example 8 (Table 4b), $\{\{\bar{S}_{T_8}\}\{\}\}$ is the derived cluster of $\mathbf{C}^{\vartheta}_{2_2}$.

Determining all derived clusters is computationally expensive however, our method only requires the derived clusters of the ultimate corpus of the serial publication (Section 3).

In subsequent sections, we frequently make use of the concept of **cover** which refers to the multiset of all possible overlaps for any given cluster $\mathbf{C}^*$. It is formally defined as follows.

Definition 8 (*Cover*)**.** Let $\mathbf{C}^* \in \mathbb{T}^*$ be a published cluster of the serial publication $\mathcal{P}^*$, the **cover** $\Omega(\mathbf{C}^*)$ returns a multiset of all the overlaps of the cluster $\mathbf{C}^*$ *i.e.* $\Omega(\mathbf{C}^*) = \{\!\{\mathbf{O}(\mathbf{C}^*, \mathbf{C}_i'^*) | \mathbf{C}_i'^* \in \mathcal{P}^*\}\!\}$.

2.3.1. Prior and posterior probability

For any transaction T^* in the published corpus $\mathbb{T}^*$, there are two levels of probabilities for linking a sensitive term s to T^*. The first is the prior probability which considers the published corpus $\mathbb{T}^*$ to be a single independent publication *i.e.* there are no other publications that can affect the probability of linking a sensitive term s to T^* in $\mathbb{T}^*$. The second is the posterior probability which considers the effect of other serial publications on $\mathbb{T}^*$. These are presented as follows.

Definition 9 (*Prior Probability*)**.** Given a private term $s \in S_{\mathcal{P}}$, the published transaction $T^* \in \mathbf{C}^*$ of $\mathbb{T}^*$, the prior probability $\alpha(T^*, s)$ of T^* w.r.t. s is the probability that its random reconstruction T^r has s (Definition 4):

$$\alpha(T^*, s) = Prob(s \in T^r) = \frac{N(s, \mathbf{C}^*)}{N(\mathbf{C}^*)} \quad (9)$$

The prior probability of an adversary reflects his/her belief that T^r has s before observing the cover $\Omega(\mathbf{C}^*)$. The posterior probability is the one that he/she has after observing $\Omega(\mathbf{C}^*)$.

Definition 10 (*Posterior Probability*)**.** Given a private term $s \in S_{\mathcal{P}}$, the published transaction $T^* \in \mathbf{C}^*$ of $\mathcal{P}^*$, the posterior probability $\beta(T^*, s)$ of T^* w.r.t. s is the probability that its random reconstruction T^r has s given the cover $\Omega(\mathbf{C}^*)$:

$$\beta(T^*, s) = Prob(s \in T^r | \Omega(\mathbf{C}^*)) \tag{10}$$

By Bayes' rule;

$$Prob(s \in T^r | \Omega(\mathbf{C}^*)) = \frac{Prob(s \in T^r) \cdot Prob(\Omega(\mathbf{C}^*) | s \in T^r)}{Prob(\Omega(\mathbf{C}^*))} \tag{11}$$

In Formula (11), $Prob(s \in T^r)$ is the prior probability of T^* (Formula (9)), $Prob(\Omega(\mathbf{C}^*))$ is the joint probability of the occurrence of the overlaps in the cover $\Omega(\mathbf{C}^*)$ and $Prob(\Omega(\mathbf{C}^*)|s \in T^r)$ is the joint probability of $\Omega(\mathbf{C}^*)$ conditioned on T^r having s. The occurrence of any overlap $\mathbf{O}_i \in \Omega(\mathbf{C}^*)$ given $s \in T^r$ is independent of the occurrence of $\mathbf{O}_j \in \Omega(\mathbf{C}^*)$ $(i \neq j)$ so the *conditional independence assumption* applies. By *normalisation* [34] Formula (11) becomes Formula (12). The denominator is the normalisation factor since $Prob(\Omega(\mathbf{C}^*)) = Prob(s \in T^r) \cdot Prob(\Omega(\mathbf{C}^*)|s \in T^r) + Prob(s \notin T^r) \cdot Prob(\Omega(\mathbf{C}^*)|s \notin T^r)$. The probabilities $Prob(\mathbf{O}|s \in T^r)$ and $Prob(\mathbf{O}|s \notin T)$ are calculated by *combinatorics* in Formulae (13) and (14).

$$Prob(s \in T^r | \Omega(\mathbf{C}^*)) = \frac{Prob(s \in T^r) \cdot \prod_{\mathbf{o}_i \in \Omega(\mathbf{C}^*)} Prob(\mathbf{O}_i | s \in T^r)}{Prob(s \in T^r) \cdot \prod_{\mathbf{o}_i \in \Omega(\mathbf{C}^*)} Prob(\mathbf{O}_i | s \in T^r) + Prob(s \notin T^r) \cdot \prod_{\mathbf{o}_i \in \Omega(\mathbf{C}^*)} Prob(\mathbf{O}_i | s \notin T^r)} \tag{12}$$

Formulae (13) and (14) represent *hypergeometric probabilities* for the selection of the overlap $\mathbf{O}$ from the cluster $\mathbf{C}^*$ containing T^*; $\binom{A}{B}$ is a combination function; and z offsets a double selection if $\bar{S}_{T^r} \in \mathbf{O}^{\bar{S}}$ since $s \in T^r$ or $s \notin T^r$ is given.

$$Prob(\mathbf{O}|s \in T^r) = \frac{\sum_{r=r_1}^{r_2} \binom{N(s,\mathbf{C}^*)-1}{r-z} \cdot \binom{N(\bar{s},\mathbf{C}^*)}{N(\mathbf{O})-r}}{\binom{N(\mathbf{C}^*)-1}{N(\mathbf{O})-z}} \tag{13}$$

$$Prob(\mathbf{O}|s \notin T^r) = \frac{\sum_{r=r_1}^{r_2} \binom{N(s,\mathbf{C}^*)}{r} \cdot \binom{N(\bar{s},\mathbf{C}^*)-1}{N(\mathbf{O})-r-z}}{\binom{N(\mathbf{C}^*)-1}{N(\mathbf{O})-z}} \tag{14}$$

$z = 1$ *if* $\bar{S}_{T^r}$ *is in* $\mathbf{O}^{\bar{S}}$ (*of* $\mathbf{O}$)*; else* $z = 0$.

In Formula (13), $\binom{N(\bar{s},\mathbf{C}^*)}{N(\mathbf{O})-r}$ is the number of ways in which $N(\mathbf{O})-r$ $\bar{s}$-transactions can be selected from $N(\bar{s}, \mathbf{C}^*)$. $\binom{N(s,\mathbf{C}^*)-1}{r-z}$ is the number of ways in which $r-z$ s-transactions can be selected from $N(s, \mathbf{C}^*)-1$ ("-1" because we have assumed $s \in T^r$). $\binom{N(\mathbf{C}^*)-1}{N(\mathbf{O})-z}$ gives all possible ways of selecting $N(\mathbf{O}) - z$ transactions from $N(\mathbf{C}^*) - 1$ ("-1" because T^r is no longer available for selection). With the same intuition Formula (14) can be explained.

Example 9. The prior and posterior probability for T_1^* (Laura) is calculated as follows. From $\mathbf{C}_{1_1}^{\vartheta}$ (Table 4a), $\alpha(T_1^*, HIV) = Prob(HIV \in T_1^r) = 1/2$ (Definition 9).

From Example 7 and Definition 8, the cover of $\mathbf{C}_{1_1}^{\vartheta}$ is $\{\mathbf{O}(\{\bar{S}_{T_2}\}, \{[0, 1]_{cancer}, [0, 0]_{HIV}\})\}$. The probability $Prob(\mathbf{O}|HIV \in T_1^r)$ is 1 (Formula (13)). In Formula (13), given $N(\mathbf{O}) = 1$, $r = r_1 = r_2 = 0$, $z = 0$, $N(\overline{HIV}, \mathbf{C}_{1_1}^{\vartheta}) = 1$, $N(HIV, \mathbf{C}_{1_1}^{\vartheta}) = 1$ and $N(\mathbf{C}_{1_1}^{\vartheta}) = 2$; in the numerator there is only 1 way to select $N(\mathbf{O})-r$ $\overline{HIV}$-transactions from $N(\overline{HIV}, \mathbf{C}_{1_1}^{\vartheta})$ and $r-z$ HIV-transactions from $N(HIV, \mathbf{C}_{1_1}^{\vartheta}) - 1$; in the denominator, there is only 1 way to select $N(\mathbf{O}) - z$ transactions from $N(\mathbf{C}_{1_1}^{\vartheta}) - 1$. With similar explanation, the probability $Prob(\mathbf{O}|HIV \notin T_1^r)$ is also calculated to be 0 (Formula (14)). With $Prob(HIV \in T_1^r)$, $Prob(\mathbf{O}|HIV \in T_1^r)$ and $Prob(\mathbf{O}|HIV \notin T_1^r)$ calculated, Formula (12) evaluates to $\frac{(1/2)\cdot 1}{(1/2)\cdot 1+(1/2)\cdot 0} = 1$, and $\beta(T_1^*, HIV) = 1$ as expected. Similarly, $\beta(T_1^*, cancer)$ is calculated to be $= 1/2$.

The following lemma summarises the relationship between the overlaps and the posterior probability.

Lemma 1. *Given a transaction* $T^* \in \mathbf{C}^*$ *of* $\mathcal{P}^*$*, the private term* $s \in S_{\mathcal{P}}$ *and its cover* $\Omega(\mathbf{C}^*)$*, the following holds:*

1. *Each overlap* $\mathbf{O}_i \in \Omega(\mathbf{C}^*)$ *is a subset of* $\mathbf{C}^*$ *and derives an inference about the private terms* $s \in S_{\mathcal{P}}$ *of* T^**.*
2. *The overall inference about the private terms* $s \in S_{\mathcal{P}}$ *of* T^* i.e. $\beta(T^*, s)$ *depends on the collective inferences from all the overlaps* $\mathbf{O}_i \in \Omega(\mathbf{C}^*)$*.*

Proof. See Appendix for proof. □

2.3.2. Global composition

When $\mathbb{T}^*$ is published, an adversary can trivially consider $\mathbb{T}^*$ as a single cluster in conducting an attack. For example, when Table 4a is wholly compared with Table 4b, we notice that $\bar{S}_{T_1}$, $\bar{S}_{T_5}$, $\bar{S}_{T_6}$, and the private term *herpes* are in Table 4a but not in Table 4b. The adversary concludes that *herpes* must be associated with the non-private sets $\bar{S}_{T_1}$, $\bar{S}_{T_5}$ and $\bar{S}_{T_6}$ which causes a knowledge gain from the prior probability of 1/6 to a new posterior probability of 1/3. To capture this *global composition*, we assume $\mathbb{T}^*$ to be partitioned into a single cluster $\mathbf{C}^*$. We set $\mathbf{C}^* = \{\mathbf{C}^{\bar{S}}, C_C^S, C_{\bar{o}l}^S\}$ to $\mathbb{T}^* = \{\mathbb{T}^{\bar{S}}, T_{\uplus}^S, \{\}\}$ in the Formulae in Section 2.3.1; where $\mathbb{T}^{\bar{S}}$ is a multiset of all the non-sensitive sets in $\mathbb{T}^*$ and $T_{\uplus}^S$ is the multiset of all sensitive terms in $\mathbb{T}^*$. Together, the maximum posterior probability of the global and cluster composition determines the **serial risk** of a published transaction T^*. This is also illustrated by Example 10.

2.4. Privacy guarantee & problem definition

Definition 11 (*Serial s-Risk*)**.** Given a private term $s \in S_{\mathcal{P}}$, a published transaction $T^* \in \mathbf{C}^*$ of $\mathcal{P}^*$ and its cover $\Omega(\mathbf{C}^*)$, the **serial risk** $\gamma_s^{ser.}$ of T^* w.r.t. the sensitive term s is the ratio of the posterior probability $\beta(T^*, s)$ to the population rate:

$$\gamma_s^{ser.}(T^*) = \frac{\beta(T^*, s)}{ss(s, \mathbb{T})} \tag{15}$$

Definition 12 (*Serially Preserving*)**.** Given a published transaction $T^* \in \mathbf{C}^*$ of $\mathcal{P}^*$, and a user defined threshold $r_{th} \in [1, N(\mathbb{T})]$, T^* is serially preserving if $\gamma_s^{ser.}(T^*) \leq r_{th}$ for every private term $s \in S_{\mathcal{P}}$. $\mathcal{P}^*$ is serially preserving if every transaction $T^* \in \mathcal{P}^*$ is serially preserving.

Serially preserving means the s-preserving guarantee of Definition 3 is always preserved.

Example 10. From Example 9, $\beta(T_1^*, HIV) = 1$ and the serial HIV-risk $\gamma_{HIV}^{ser.}(T_1^*)$ of transaction T_1^* is $\frac{1}{1/3} = 3$. For $r_{th} = 2$, T_1^* is not serially preserving since $3 > 2$. Notice that, due to global composition, $\gamma_{Herpes}^{ser.}(T_1^*) = 2$ although it is 0 when only $\mathbf{C}_{1_1}^{\vartheta}$ and $\mathbf{C}_{1_2}^{\vartheta}$ (Table 4) are considered.

Our problem now is to find a suitable publication mechanism π that satisfies this privacy guarantee (Definition 12).

Definition 13 (*Problem Definition*)**.** Let $\mathcal{P}^* = [\mathbb{T}_1^*, \ldots, \mathbb{T}_m^*]$ be serially preserving. Given the anonymised corpus $\mathbb{T}_{m+1}^{\vartheta}$, the problem that this paper seeks to address is to find a publication mechanism π such that the publication $\mathcal{P}^* + [\pi(\mathbb{T}_{m+1}^{\vartheta})]$ remains serially preserving.

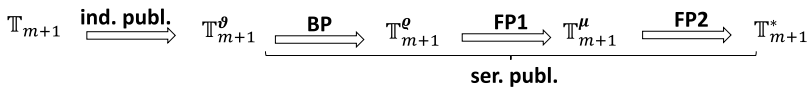


Fig. 5. Serial publication process.

In this section we developed a privacy guarantee for the serial publication of transactional data. We first presented a privacy guarantee for the single independent publication scenario. Then, we developed this guarantee further by considering the risk of privacy disclosures based on the posterior knowledge of the adversary in the serial publication setting. The posterior knowledge of the adversary was defined as the posterior probability of a transaction having a sensitive term after the adversary observes other publications in the serial publication. In calculating this probability, we provided some theoretical reasoning which made use of combinatorics and the Bayes' theorem to maximise the adversary's knowledge. We further analysed the risk by considering the possibility of global compositions, where the adversary can make use the whole of each published corpus to enhance his/her posterior knowledge about a transaction. Finally, we formally presented the problem this paper seeks to address.

In the following, we present our proposed publication mechanism which addresses the privacy preserving serial publication problem (Definition 13).

3. Framework of the solution

In this section, we present our solution *Sanony* to the serial publication problem (Definition 13). While composition attacks can occur only in the presence of overlaps, it is not desirable to remove such overlapping transactions as this would lead to high utility losses. Instead, a good solution should render the overlaps ineffective for any meaningful inference without the prohibitive loss of utility. This is the basis for all existing solutions treating this problem. However, the special nature of our problem in terms of its stronger privacy requirement (Definition 12) and the schema-less nature of our input transactional data requires a tailored solution.

In our solution, we establish theoretically that the presence of counterfeits reduces the ability of the adversary to increase his/her knowledge about the transactions. We then develop a two step approach to protect the privacy of previously published corpora as well as current and future publications. We conduct further analysis on our solution and ensure that it is fool-proof against minimality attacks. Finally we present our algorithm *Sanony* and its analysis. Our solution is described as follows.

Given a serially preserving corpora $\mathcal{P}^*$ (previously published) and the anonymised corpus $\mathbb{T}^{\vartheta}_{m+1}$ (newly anonymised to be published corpus), *Sanony* proceeds in two main steps of perturbation as follows (Fig. 5).

- *Backward Perturbation* (*BP*) adds counterfeits to each cluster $\mathbf{C}^{\vartheta}$ of $\mathbb{T}^{\vartheta}_{m+1}$ to alter its cover $\Omega(\mathbf{C}^{\vartheta})$ so that the inferences that can be made from $\Omega(\mathbf{C}^{\vartheta})$ is reduced. This removes the risk of composition attacks from linking $\mathbf{C}^{\vartheta}$ to the transactions of the previously published corpora.
- *Forward Perturbation* (*FP*1, *FP*2) adds further counterfeits to the clusters to be published after *BP* to ensure that their derived clusters (Definition 7) cannot lead to transitive composition attacks; and each transaction remains serially preserving. This, in addition to BP removes the risk of composition attacks to the transactions of the newly anonymised corpus to be published.

Although counterfeits are required in the two stages, the number of counterfeits added in practice is always observed to be a small percentage. From our experiments (Section 4), the maximal number of counterfeits added constitutes less than 4% of the dataset while the average is less than 1%.

3.1. Backward perturbation

Given a serially preserving corpus $\mathcal{P}^* = [\mathbb{T}^*_1, \ldots, \mathbb{T}^*_m]$, all transactions $T^* \in \mathcal{P}^*$ are guaranteed to be serially preserving (Definition 12). When a newly anonymised corpus $\mathbb{T}^{\vartheta}_{m+1}$ is published, the overlap $\mathbf{O}(\mathbf{C}^*, \mathbf{C}^{\vartheta})$ ($\mathbf{C}^* \in \mathcal{P}^*, \mathbf{C}^{\vartheta} \in \mathbb{T}^{\vartheta}_{m+1}$) may cause an update to the posterior probabilities of the transactions $T^* \in \mathcal{P}^*$. Consequently, $T^* \in \mathcal{P}^*$ may no longer be serially preserving (Section 2.3). Compelling the updated posterior probability $\beta(T^*, s)$ to tend back to its prior probability $\alpha(T^*, s)$ when $\mathbb{T}^{\vartheta}_{m+1}$ is published will ensure that $T^* \in \mathcal{P}^*$ remains serially preserving.

Given an overlap $\mathbf{O}(\mathbf{C}^*, \mathbf{C}^{\vartheta})$ of cluster $\mathbf{C}^*$ in the published corpora and cluster $\mathbf{C}^{\vartheta}$ in the anonymised corpus to be published, this section shows (1) that the posterior probability $\beta(T^*, s)$ reduces as the interval of the range $[r_1, r_2]_s$ of $\mathbf{O}$ increases ($T^* \in \mathbf{C}^*$); and (2) how counterfeits should be computed and added to $\mathbf{C}^{\vartheta}$ to transform the range $[r_1, r_2]_s$ such that the posterior probability $\beta(T^*, s)$ approaches the prior probability $\alpha(T^*, s)$. The first point above is characterised by Lemma 2, while the second point is characterised by Lemmas 3 and 4 in the following.

Lemma 2. *Let $\mathbf{O}(\mathbf{C}^*, \mathbf{C}^{\vartheta})$ be an overlap where $\mathbf{C}^*$ is in $\mathcal{P}^*$ and $\mathbf{C}^{\vartheta}$ is in $\mathbb{T}^{\vartheta}_{m+1}$. Given $s \in S_{\mathcal{P}}$ and its range $[r_1, r_2]_s$ in $\mathbf{O}$, the posterior probability $\beta(T^*, s)$ approaches the prior probability $\alpha(T^*, s)$ as the interval of $[r_1, r_2]_s$ increases.*

Proof. See Appendix for proof. □

In the following, Lemma 3 prescribes the ideal range $[r_1, r_2]_s$ for an overlap $\mathbf{O}(\mathbf{C}^*, \mathbf{C}^{\vartheta})$ such that for $T^* \in \mathbf{C}^*$, $\beta(T^*, s) = \alpha(T^*, s)$ is true.

Lemma 3. *Let $\mathbf{O}(\mathbf{C}^*, \mathbf{C}^{\vartheta})$ be an overlap where $\mathbf{C}^*$ is in $\mathcal{P}^*$ and $\mathbf{C}^{\vartheta}$ is in $\mathbb{T}^{\vartheta}_{m+1}$. Given the private term $s \in S_{\mathcal{P}}$, $\beta(T^*, s) = \alpha(T^*, s)$ ($T^* \in \mathbf{C}^*$) if the s-range $[f_{1_{\mathbf{C}^*}}, f_{2_{\mathbf{C}^*}}]_s$ (Definition 5) lies in the range $[r_1, r_2]_s$ of $\mathbf{O}$.*

Proof. See Appendix for proof. □

If the range of an overlap $\mathbf{O}$ satisfies Lemma 3, we say it is **safe**. Safe overlaps can be achieved by making use of counterfeit transactions, first we define what a **counterfeit transaction** is as follows.

Definition 14 (*Counterfeit Transaction*)**.** Let $\mathbf{C}^{\vartheta}$ be an anonymised cluster, $\mathbf{C}^{\bar{S}}$ its non-private set and $C^{\bar{S}}_{\uplus}$ a multi-set of its non-private terms, a counterfeit transaction T_c is a term set containing a random sample of the minimal number of terms from $C^{\bar{S}}_{\uplus}$ such that T_c is not identical to any transaction in $\mathbf{C}^{\bar{S}}$.

When a private term s is added to T_c, T_c is called an **s-counterfeit** else it is an **$\bar{s}$-counterfeit**. The terms of a counterfeit transaction are generated from the terms within the cluster rather than arbitrarily to minimise the utility loss impact from the presence of counterfeits. By ensuring that T_c is not identical to any of the existing transactions, a reduction in posterior probability when counterfeits are added is guaranteed (proof of Lemma 4). Where it is not possible to obtain a T_c that is not identical to any transaction

Table 6
Backward perturbation.

(a) $\mathbb{T}_1^{\varrho}$

$\mathbf{C}^{\bar{S}}$	C_c^S	$C_{o^J}^S$
$\mathbf{C}_{1_1}^{\varrho}$		
$\bar{S}_{T_1}$	*HIV*,	
$\bar{S}_{T_2}$	*cancer*	

(b) $\mathbb{T}_2^{\varrho}$

$\mathbf{C}^{\bar{S}}$	C_c^S	$C_{o^J}^S$
$\mathbf{C}_{1_2}^{\varrho}$		
$\bar{S}_{T_2}$	*cancer*,	
$\bar{S}_{T_7}$	*HIV*$^{\diamond}$	
$T_c^{\diamond}$		

(c) $\mathbb{T}_3^{\varrho}$

$\mathbf{C}^{\bar{S}}$	C_c^S	$C_{o^J}^S$
$\mathbf{C}_{1_3}^{\varrho}$		
$\bar{S}_{T_2}$	*cancer*,	
$\bar{S}_{T_7}$	*HIV*	
$\bar{S}_{T_{10}}$		

in $\mathbf{C}^{\bar{S}}$ *i.e.* when the cluster contains only one transaction, T_c is sampled from $T_{\uplus}^{\bar{S}}$ the multiset of all non-private terms of the corpus. In the extreme case where neither of the above is possible then T_c is sampled from $T_{1_{\uplus}}^{\bar{S}} \biguplus, \ldots, \biguplus T_{m_{\uplus}}^{\bar{S}}$, the multiset union of all non-private terms in the corpora.

The following Lemma 4 defines how the counterfeits required to make the overlap **O** safe should be calculated.

Lemma 4. *Let* $\mathbf{O}(\mathbf{C}^*, \mathbf{C}^{\vartheta})$ *be an overlap where* $\mathbf{C}^*$ *is in* $\mathcal{P}^*$ *and* $\mathbf{C}^{\vartheta}$ *is in* $\mathbb{T}_{m+1}^{\vartheta}$. *Given* $s \in S_{\mathcal{P}}$, *its range* $[r_1, r_2]_s$ *in* **O** *and the s-range* $[f_{1_{\mathbf{C}^*}}, f_{2_{\mathbf{C}^*}}]_s$ *(Definition 5),* $\mathbf{C}^{\vartheta}$ *requires at least* $h_{\bar{s}}$ *number of* $\bar{s}$*-counterfeits and* h_s *number of s-counterfeits to make* **O** *safe:*

$$h_{\bar{s}} = h_{\bar{s}}(\mathbf{O}) = (r_1 - f_{1_{\mathbf{C}^*}}) \qquad h_s = h_s(\mathbf{O}) = (f_{2_{\mathbf{C}^*}} - r_2) \tag{16}$$

$h_{\bar{s}}$ *and* h_s *are both non-negative due to Proposition 1.*

Proof. See Appendix for proof. □

In Lemma 4, the addition of more counterfeits only maintains $r_1' \leq f_{1_{\mathbf{C}^*}}$ or $r_2' \geq f_{2_{\mathbf{C}^*}}$, and Lemma 3 remains satisfied.

Given $\mathbf{C}^{\vartheta}$, for each $s \in S_{\mathcal{P}}$ it suffices from the above result, to add the maximum number of counterfeits required for all overlaps in the cover $\Omega(\mathbf{C}^{\vartheta})$ to make every overlap safe. Let $h_x(\mathbf{O}_i)$ $(x \in \{s, \bar{s}\})$ be the number of x-counterfeits required to make $\mathbf{O}_i \in \Omega(\mathbf{C}^{\vartheta})$ safe w.r.t. x. The maximum number of counterfeits required for all overlaps is given by:

$$h_x(\Omega(\mathbf{C}^{\vartheta})) = \max_{\mathbf{O}_i \in \Omega(\mathbf{C}^{\vartheta})} \{h_x(\mathbf{O}_i)\} \tag{17}$$

The maximum number of counterfeits (of both s and $\bar{s}$) for all overlaps in $\Omega(\mathbf{C}^{\vartheta})$ and all private terms $S_{\mathcal{P}}$ is given by:

$$h(\Omega(\mathbf{C}^{\vartheta})) = \max_{s \in S_{\mathcal{P}}} \{h_{\bar{s}}(\Omega(\mathbf{C}^{\vartheta})) + h_s(\Omega(\mathbf{C}^{\vartheta}))\} \tag{18}$$

Definition 15 (*Backward Perturbation*)**.** Given an anonymised cluster $\mathbf{C}^{\vartheta} = \{\mathbf{C}^{\bar{S}}, C_c^S, C_{o^J}^S\}$ to be published, and its cover $\Omega(\mathbf{C}^{\vartheta})$, **backward perturbation** *BP* adds $h(\Omega(\mathbf{C}^{\vartheta}))$ number of counterfeits (Definition 14) to $\mathbf{C}^{\bar{S}}$; and $h_s(\Omega(\mathbf{C}^{\vartheta}))$ copies of s to C_c^S for each private term $s \in S_{\mathcal{P}}$.

After BP (Fig. 5) $\mathbf{C}^{\vartheta}$ becomes $\mathbf{C}^{\varrho}$, it is illustrated by Example 11.

Example 11. $\mathbf{C}_{1_1}^{\varrho}, \mathbf{C}_{1_2}^{\varrho}$ and $\mathbf{C}_{1_3}^{\varrho}$ (Table 6) is the backward perturbation on $\mathbf{C}_{1_1}^{\vartheta}, \mathbf{C}_{1_2}^{\vartheta}$ and $\mathbf{C}_{1_3}^{\vartheta}$ (Table 4) respectively. In the table, the actual terms of the non-private sets are excluded for brevity. By Definition 15, one counterfeit and a copy of *HIV* needs to be added to $\mathbf{C}^{\bar{S}}$ and C_c^S of $\mathbf{C}_{1_2}^{\vartheta}$ respectively (marked $\diamond$ for illustration only). No counterfeits are required to be added to $\mathbf{C}_{1_3}^{\vartheta}$. The overlap between $\mathbf{C}_{1_1}^{\varrho}$ and $\mathbf{C}_{1_2}^{\varrho}$ is now $\{\{\bar{S}_{T_2}\}, \{[0,1]_{cancer}, [0,1]_{HIV}\}\}$ and both ranges satisfies Lemma 3. Also, the overlap between $\mathbf{C}_{1_2}^{\varrho}$ and $\mathbf{C}_{1_3}^{\varrho}$ is $\{\{\bar{S}_{T_2}, \bar{S}_{T_7}\}, \{[0,1]_{cancer}, [0,1]_{HIV}\}\}$ which is also safe.

In global composition, the overlap is between two corpuses as a whole, therefore the counterfeit is added to the non-private set of the cluster that is most similar to the counterfeit (*e.g. Jaccard similarity* or *cosine similarity*) and the private terms are added into the global bag.

After BP every transaction T^* in $\mathcal{P}^*$ remains serially preserving when $[\mathbb{T}_{m+1}^{\varrho}]$ is published due to Lemma 3.

3.2. Forward perturbation

Following the s-preserving definition (Definition 3), every transaction $T^{\vartheta} \in \mathbb{T}_{m+1}^{\vartheta}$ to be published is s-preserving, and generally speaking is serially preserving when $\mathbb{T}_{m+1}^{\vartheta}$ is an independent publication. However, when $\mathbb{T}_{m+1}^{\vartheta}$ is to be published as part of serial publication $\mathcal{P}^* = [\mathbb{T}_1^*, \ldots, \mathbb{T}_m^*]$, $\mathbb{T}_{m+1}^{\vartheta}$ may not be serially preserving. It may also be used to form transitive composition attacks (Section 2.3), and therefore requires some form of perturbation.

Forward perturbation proceeds in two operational steps, FP1 and FP2 (Fig. 5). In this section, we first present FP1 which removes the possibility of transitive composition attacks by developing (1) how the range of $s \in S_{\mathcal{P}}$ in the derived clusters should be calculated when the number of counterfeits are published (Proposition 2); and (2) how the counterfeits required in FP1 should be computed (Lemma 5). Second, we present FP2 which ensures that after FP1 the transactions (to be published) remain serially preserving by adding some counterfeits. Together, FP1 and FP2 finalise the publication process of $\mathbb{T}_{m+1}^{\varrho}$ to $\mathbb{T}_{m+1}^{*}$.

In the following, the superscript ϱ denotes the result of backward perturbation (Section 3.1) and μ denotes the intermediate result of FP1.

Proposition 2. *Given a private term* $s \in S_{\mathcal{P}}$ *and the overlap* $\mathbf{O}(\mathbf{C}^*, \mathbf{C}^{\varrho})$ *(shortened* **O***) where* $\mathbf{C}^*$ *is in* $\mathcal{P}^*$ *and* $\mathbf{C}^{\varrho}$ *is in* $\mathbb{T}_{m+1}^{\varrho}$, *let* $\mathbf{d}^{\bar{S}} = \mathbf{C}^{\bar{S}} \setminus \mathbf{O}^{\bar{S}}$ $(\mathbf{C}^{\bar{S}} \in \mathbf{C}^{\varrho})$ *be the non-overlapping transactions of* $\mathbf{C}^{\varrho}$ *and* x_d *be the number of counterfeits added to* $\mathbf{C}^{\varrho}$. *The number of copies of s shared by* $\mathbf{d}^{\bar{S}}$ *is in* $[r_{1_d}, r_{2_d}]_s$:

$$r_{1_d} = \max\{r_{1_f} - x_d, 0\} \qquad r_{2_d} = \min\{(N(\mathbf{C}^{\varrho}) - x_d - r_1), r_{2_f}\} \tag{19}$$

$[r_{1_f}, r_{2_f}]_s$ *is calculated by Formula (8) and* $[r_1, r_2]_s$ *is the range of s in* **O**.

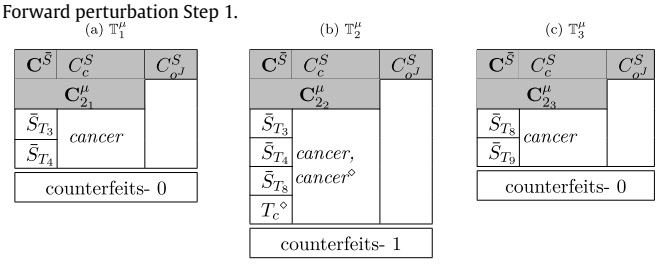
r_{1_d} is developed as follows. $N(s, \mathbf{C}^{\varrho})$ is the number of copies of s in $\mathbf{C}^{\varrho}$. In the extreme case, the overlap **O** has r_2 copies of s so the number of copies of s available for $\mathbf{d}^{\bar{S}}$ is $r_{1_f} = N(s, \mathbf{C}^{\varrho}) - r_2$(Formula (8)). For the lower bound r_{1_d} the counterfeits added x_d must be assumed to be s-counterfeits, so only $r_{1_f} - x_d$ number of true copies of s remain to be shared by $\mathbf{d}^{\bar{S}}$ and $\max\{r_{1_f} - x_d, 0\}$ in the formula ensures non-negative values.

r_{2_d} is developed as follows. $N(\mathbf{C}^{\varrho})$ is the number of transactions in $\mathbf{C}^{\varrho}$. Of these, **O** has r_1 copies of s in the extreme case and $N(\mathbf{C}^{\varrho}) - r_1$ possible number of copies of s remain. For the upper bound r_{2_d} the counterfeits added x_d must be assumed to be $\bar{s}$-counterfeits so $N(\mathbf{C}^{\varrho}) - r_1 - x_d$ copies of s remain to be shared by $\mathbf{d}^{\bar{S}}$. However, there are $N(s, \mathbf{C}^{\varrho})$ copies of s in $\mathbf{C}^{\varrho}$, if **O** has r_1 copies of s then $r_{2_f} = N(s, \mathbf{C}^{\varrho}) - r_1$ remain to be shared by $\mathbf{d}^{\bar{S}}$. We compare the two values to get the minimum because it is possible that $(N(\mathbf{C}^{\varrho}) - x_d - r_1) < r_{2_f}$ or $(N(\mathbf{C}^{\varrho}) - x_d - r_1) > r_{2_f}$, so $\min\{(N(\mathbf{C}^{\varrho}) - x_d - r_1), r_{2_f}\}$ gives the actual maximum number of s available for $\mathbf{d}^{\bar{S}}$.

The range $[r_{1_f}, r_{2_f}]_s$ of possible copies of s in the derived cluster $\mathbf{d}(\mathbf{C}^{\varrho}, \mathbf{O})$ (Definition 7) becomes $[r_{1_d}, r_{2_d}]_s$ (Proposition 2). For instance, in $\mathbf{C}^* = \{\{\!\{\begin{smallmatrix}A\\B\end{smallmatrix}\}\!\}, \{\!\{s_1\}\!\}, \{\!\{\}\!\}\}$ and $\mathbf{C}^{\varrho} = \{\{\!\{\begin{smallmatrix}A\\T_c\end{smallmatrix}\}\!\}, \{\!\{s_1\}\!\}, \{\!\{\}\!\}\}$, the overlap **O** is $\{A\}$, $\{[0,1]_{s_1}\}$ and $x_d = 1$. $[r_{1_f}, r_{2_f}]_{s_1}$ is $[0,1]_{s_1}$ (Definition 7) and $[r_{1_d}, r_{2_d}]_{s_1}$ is also $[0,1]_{s_1}$ by Proposition 2 and the derived cluster $\mathbf{d}(\mathbf{C}^{\varrho}, \mathbf{O})$ is $\{\{T_c\}, \{[0,1]_{s_1}\}\}$.

In transitive composition attacks, the adversary wishes to propagate the knowledge obtained from a derived cluster $\mathbf{d}$ to infer the private terms of some non-private set $\bar{S}_T$. When the number of counterfeits are published, (s)he can use at most $N(\mathbf{d}) - x_d$

Table 7
Forward perturbation Step 1.



"true" non-overlapping transactions for the attack. Our aim is to ensure that for any $N(\mathbf{d}) - x_d$ non-overlapping transactions in $\mathbf{d}$, there can be no inference about other non-private sets. Lemma 5 determines the counterfeits to be added to $\mathbf{C}^{\varrho}$ via FP1 to prevent such transitive composition attacks. We denote the true derived cluster of $\mathbf{d}(\mathbf{C}^{\varrho}, \mathbf{O})$, prior to the addition of counterfeits, by $\mathbf{d}_o = \mathbf{d}(\mathbf{C}^{\vartheta}, \mathbf{O})$ for the lemma as follows.

Lemma 5. *Let $\mathbf{O}(\mathbf{C}^*, \mathbf{C}^{\varrho})$ be an overlap where $\mathbf{C}^*$ is in $\mathcal{P}^*$ and $\mathbf{C}^{\varrho}$ is in $\mathbb{T}^{\varrho}_{m+1}$, also let $\mathbf{d}(\mathbf{C}^{\varrho}, \mathbf{O})$ be a derived cluster and $\mathbf{d}_o$ its true derived cluster. Given $s \in S_{\mathcal{P}}$ and its range $[r_{1_d}, r_{2_d}]_s$ in $\mathbf{d}$ (Proposition 2), the addition of $d_{\bar{s}}$ $\bar{s}$-counterfeits and d_s s-counterfeits to $\mathbf{C}^{\varrho}$, guarantees there cannot be any transitive attack via $\mathbf{d}$ w.r.t. s:*

$$\begin{aligned} d_{\bar{s}} &= d_{\bar{s}}(\mathbf{O}) = N(s, \mathbf{C}^{\varrho}) - r_2 - x_d \\ d_s &= d_s(\mathbf{O}) = N(\mathbf{d}_o) + r_1 - N(s, \mathbf{C}^{\varrho}) \end{aligned} \tag{20}$$

x_d is the number of counterfeits added after BP. $[r_1, r_2]_s$ is the range of s in $\mathbf{O}$.

Proof. See Appendix for proof. □

Similar to Formula (17) and (18) also let $d_x(\Omega(\mathbf{C}^{\varrho}))$ and $d(\Omega(\mathbf{C}^{\varrho}))$ be the maximum number of x-counterfeits and the overall maximum number of counterfeits respectively required for all overlaps $\Omega(\mathbf{C}^{\varrho})$. **FP1** is defined as follows.

Definition 16 (*FP1*)**.** Given a backward perturbed cluster $\mathbf{C}^{\varrho} = \{\mathbf{C}^{\bar{S}}, C^S_c, C^S_{o'}\}$, and its cover $\Omega(\mathbf{C}^{\varrho})$, **FP1** adds $d(\Omega(\mathbf{C}^{\vartheta}))$ number of counterfeits (Definition 14) to $\mathbf{C}^{\bar{S}}$; and $d_s(\Omega(\mathbf{C}^{\vartheta}))$ copies of s to C^S_c for each $s \in S_{\mathcal{P}}$.

After FP1, the cluster $\mathbf{C}^{\varrho}$ becomes $\mathbf{C}^{\mu}$ and it's transactions are no longer at risk of transitive composition attacks due to Lemma 5. The transactions T^* of the previously published corpora $\mathcal{P}^*$ are also safe from transitive composition attacks since they follow the same publication scheme.

Example 12. $\mathbf{C}^{\mu}_{2_1}$, $\mathbf{C}^{\mu}_{2_2}$ and $\mathbf{C}^{\mu}_{2_3}$ (Table 7) is the result of BP and FP1 on $\mathbf{C}^{\vartheta}_{2_1}$, $\mathbf{C}^{\vartheta}_{2_2}$ and $\mathbf{C}^{\vartheta}_{2_3}$ (Table 4) respectively. By Lemma 5, $\mathbf{C}^{\mu}_{2_2}$ needs 1 cancer-counterfeit (marked $\diamond$ for illustration only). The overlap of $\mathbf{C}^{\mu}_{2_2}$ and $\mathbf{C}^{\mu}_{2_3}$ is $\{\{\bar{S}_{T_8}\}, \{[0, 1]_{cancer}\}\}$ and there can be no transitive composition attack as previously seen in Example 8.

After BP and FP1 the number of counterfeits added are published along with the cluster. The aim of publishing the counterfeits is two fold (1) to add some perceived utility to the clusters; and (2) to remove the possibility of further transitive attacks due to the addition of counterfeits after FP1.

The publication of the number of counterfeits prevents adversary from confidently using the derived clusters after the addition of counterfeits to conduct further transitive composition attacks. The known number of counterfeits 'forces' the adversary to consider these counterfeits in his/her attack. For instance, if the number of counterfeits in $\mathbf{C}^{\mu}_{2_2}$ (Table 7b) is not known, the adversary takes the derived cluster $\{\bar{S}_{T_8}, \bar{S}_{T_c}\}, \{(1, 1)_{cancer}\}$ of $\mathbf{C}^{\mu}_{2_2}$ in good faith. Suppose there are 3 transactions instead of 2 in $\mathbf{C}^{\mu}_{2_3}$ *i.e.* $\mathbf{C}^{\mu}_{2_3} = \{\{\begin{smallmatrix}\bar{S}_{T_8}, \bar{S}_{T_9} \\ \bar{S}_{T_{11}}\end{smallmatrix}\}, \{\begin{smallmatrix}cancer \\ cancer\end{smallmatrix}\}, \{\}\}$ and $\bar{S}_{T_{11}}$ is identical to $T_c{}^{\diamond}$. (S)he identifies the private term of $\bar{S}_{T_9}$ as *cancer* by using the derived cluster $\{\bar{S}_{T_8}, \bar{S}_{T_c}\}, \{(1, 1)_{cancer}\}$. The published number of counterfeits ironically introduces the uncertainty required to forestall such privacy disclosure.

At the same time, publishing the number of counterfeits does not lead to any disclosure risk, since the s or $\bar{s}$ status of the counterfeits is not specified; and it cannot be derived from the knowledge of the publication mechanism because the process of adding s or $\bar{s}$ counterfeits is symmetrical. For instance, $\mathbf{C}^* = \{\{\begin{smallmatrix}A \\ B\end{smallmatrix}\}, \{s_1\}, \{\}\}$ and $\mathbf{C}^{\vartheta} = \{\{A\}, \{s_1\}, \{\}\}$ differ only on B. By Definition 15 one counterfeit must be added to $\mathbf{C}^{\vartheta}$ to become $\{\{\begin{smallmatrix}A \\ T_c\end{smallmatrix}\}, \{s_1\}, \{\}\}$. Even when the adversary knows T_c is a counterfeit, (s)he does not know whether T_c is s or $\bar{s}$, consequently (s)he cannot breach the privacy of A and Lemma 3 is still satisfied.

The addition of counterfeits in backward perturbation *BP* and forward perturbation FP1 has two unlikable effects; (1) the population rates before and after BP and FP1 are not equal *i.e.* $ss(s, \mathbb{T}^{\vartheta}_{m+1}) \neq ss(s, \mathbb{T}^{\mu}_{m+1})$ due to the addition of counterfeits. $ss(s, \mathbb{T}^{\mu}_{m+1})$ becomes the new population rate; (2) the transactions in $\mathbb{T}^{\mu}_{m+1}$ may no longer be s-preserving (Definition 3). **FP2**, defined in the following, ensures that every transaction $T^{\mu} \in \mathbb{T}^{\mu}_{m+1}$ becomes serially preserving.

Definition 17 (*FP2*)**.** Let $\mathbb{T}^{\mu}_{m+1}$ be the result of applying FP1 to $\mathbb{T}^{\varrho}_{m+1}$. Given the private term $s \in S_{\mathcal{P}}$ and the risk threshold r_{th}, if $T^{\mu} \in \mathbf{C}^{\mu}$ ($\mathbf{C}^{\mu} \in \mathbb{T}^{\mu}_{m+1}$) is found not to be serially preserving (Definition 3), FP2 is the process of adding x_c counterfeits to $\mathbf{C}^{\mu}$ to make it serially preserving.

x_c is determined by gradually adding a counterfeit to $\mathbf{C}^{\mu}$ until every transaction $T^{\mu} \in \mathbf{C}^{\mu}$ is serially preserving. This is repeated for all transactions $T^{\mu} \in \mathbb{T}^{\mu}_{m+1}$ after which $\mathbb{T}^{\mu}_{m+1}$ becomes the published version $\mathbb{T}^*_{m+1}$.

Lemma 6 shows that when $\mathbf{C}^{\mu}$ is published to $\mathbf{C}^*$ after FP2, every transaction $T^* \in \mathbf{C}^*$ is serially preserving.

Lemma 6. *Given a private term $s \in S_{\mathcal{P}}$, a riskslift threshold r_{th}, and a transaction $T^{\mu} \in \mathbf{C}^{\mu}$ of $\mathbb{T}^{\mu}_{m+1}$ such that T^{μ} is not serially preserving w.r.t. s. T^{μ} becomes serially preserving when $x_c \leq \theta_c$ $\bar{s}$-counterfeits are added to $\mathbf{C}^{\mu}$; where $\theta_c = \frac{A}{ss(s, \mathbb{T}^{\mu}) \cdot r_{th}} - B$ is the upper bound on the $\bar{s}$-counterfeits required and A/B is the posterior probability $\beta(T^{\mu}, s)$.*

Proof. See Appendix for proof. □

3.2.1. Further analysis on backward perturbation and forward perturbation

After FP2, we do not publish the number of counterfeits added in this step to avoid a minimality attack. In an extreme case (*e.g.* $\mathbb{T}^{\vartheta}_{m+1}$ is partitioned into a single cluster $\mathbf{C}^{\vartheta}$ and no counterfeits were added/needed during BP or FP1), given the published population rate $ss(s, \mathbb{T}^{\mu}_{m+1})$ after FP1, when $T^{\mu} \in \mathbf{C}^{\mu}$ is found not to be serially s-preserving further $\bar{s}$-counterfeits are added by the FP2 step. Since only $\bar{s}$-counterfeits are added in FP2, the adversary may attempt to use the number of $\bar{s}$-counterfeits added in FP2 (if provided), to reverse the process and hence breach privacy of T^{μ}.

It must be noted that the non-publication of the counterfeits added in the FP2 step does not increase the chances of a disclosure. In other words, it is not possible for a persevering adversary to recalculate the number of $\bar{s}$-counterfeits added by FP2 using the published population rate $ss(s, \mathbb{T}^{\mu}_{m+1})$ and the observed population rate $ss(s, \mathbb{T}^*_{m+1})$ when $\mathbb{T}^*_{m+1}$ is finally published. We consider the worst

case where $\mathbb{T}^*_{m+1}$ is partitioned into a single cluster again. From the observed population rate $ss(s, \mathbb{T}^*_{m+1})$, $N(\mathbb{T}^*) - ss(s, \mathbb{T}^*_{m+1}) \times N(\mathbb{T}^*)$ gives the number $\bar{s}$-transactions in $\mathbb{T}^*$. But $N(\mathbb{T}^*) - ss(s, \mathbb{T}^{\mu}_{m+1}) \times N(\mathbb{T}^*)$ using the published population rate $ss(s, \mathbb{T}^{\mu}_{m+1})$ does NOT give the number of $\bar{s}$-transactions in $\mathbb{T}^{\mu}$ (remembering that $\mathbb{T}^*$ but NOT $\mathbb{T}^{\mu}$ is published). Therefore the error introduced by publishing the pair $ss(s, \mathbb{T}^{\mu}_{m+1})$ and $\mathbb{T}^*$ allows us to mitigate minimality attacks and privacy is protected.

Backward and forward perturbation form the publication $\pi(\mathbb{T}^{\vartheta}_{m+1})$ for the serial corpus $\mathbb{T}^{\vartheta}_{m+1}$. The algorithm *Sanony* is described in the following section.

3.3. Algorithm

Sanony is shown in Algorithm 1. In the algorithm, a new corpus $\mathbb{T}_{m+1}$ to be published is first anonymised to $\mathbb{T}^{\vartheta}_{m+1}$ via *Anony* (Section 2.2). With the overlaps between the clusters of the anonymised corpus $\mathbb{T}^{\vartheta}_{m+1}$ and the previously published corpora $\mathcal{P}^*$ computed, each cluster $\mathbf{C}^{\vartheta}_i \in \mathbb{T}^{\vartheta}_{m+1}$ is backward perturbed to produce $\mathbf{C}^{\varrho}_i$ (Lines 1–3). The overlaps are recomputed and each cluster $\mathbf{C}^{\varrho}_i \in \mathbb{T}^{\varrho}_{m+1}$ is forward perturbed (FP1) (Lines 4–6) to produce $\mathbf{C}^{\mu}_i$. Finally the risk of every transaction in $\mathbb{T}^{\mu}_{m+1}$ is computed and FP2 applied (Lines 7–9). The publication $\pi(\mathbb{T}^{\vartheta}_{m+1})$ is returned as $\mathbb{T}^*_{m+1}$ which is serially preserving (Theorem 1).

Algorithm 1 Sanony (π)

Input: A serially preserving corpora $\mathcal{P}^*$; the anonymised corpus $\mathbb{T}^{\vartheta}_{m+1}$; the privacy parameter r_{th}.
Output: The serially preserving corpus $\mathbb{T}^*_{m+1}$

1: **for** every anonymised cluster $\mathbf{C}^{\vartheta}_i \in \mathbb{T}^{\vartheta}_{m+1}$ **do**
2: $\quad\mathbf{C}^{\varrho}_i \leftarrow$ Call backward perturbation BP (Definition 15).
3: **end for**
4: **for** every backward perturbed cluster $\mathbf{C}^{\varrho}_i \in \mathbb{T}^{\varrho}_{m+1}$ **do**
5: $\quad\mathbf{C}^{\mu}_i \leftarrow$ Call forward perturbation FP1 (Definition 16).
6: **end for**
7: **for all** transactions $T^{\mu}_i \in \mathbb{T}^{\mu}_{m+1}$ **do**
8: $\quad T^*_i \leftarrow$ Call forward perturbation FP2 (Definition 17).
9: **end for**
10: Return $\mathbb{T}^*_{m+1}$

The time complexity of *Sanony* is $O(Q \cdot |S_{\mathcal{P}}| + N(\mathbb{T}^{\vartheta}_{m+1}) \cdot |S_{\mathcal{P}}| \cdot \theta_c)$ where Q is the total number of overlaps $\mathbf{O}(\mathbf{C}^*, \mathbf{C}^{\vartheta})$ for all pairs of $\mathbf{C}^*$ in $\mathcal{P}^*$ and $\mathbf{C}^{\vartheta}$ in $\mathbb{T}^{\vartheta}_{m+1}$; $|S_{\mathcal{P}}|$ is the number of s terms; $N(\mathbb{T}^{\vartheta}_{m+1})$ is the number of transactions in $\mathbb{T}^{\vartheta}_{m+1}$; and θ_c is the number of counterfeits required in FP2 (worst case).

In both BP and FP1, each overlap is scanned once for every private term $s \in S_{\mathcal{P}}$ to compute the counterfeits and this complexity is $O(Q \cdot |S_{\mathcal{P}}|)$. In FP2, every transaction is scanned once for every s term. For each transaction at risk, θ_c counterfeits are added iteratively in the worst case to its cluster. Its complexity is $O(|S_{\mathcal{P}}| \cdot N(\mathbb{T}^{\vartheta}_{m+1}) \cdot \theta_c)$. Together, the complexity of the algorithm is $O(Q \cdot |S_{\mathcal{P}}| + N(\mathbb{T}^{\vartheta}_{m+1}) \cdot |S_{\mathcal{P}}| \cdot \theta_c)$. Most of this cost comes from the pairwise computation of the overlaps Q which is proportional to $N(\mathbb{T}^*) \times N(\mathbb{T}^{\vartheta}_{m+1})$ ($\mathbb{T}^* \in \mathcal{P}^*$) for each pair of corpora. In our experiments we demonstrate an extreme case where $\mathbb{T}^{\vartheta}_{m+1}$ and $\mathbb{T}^*$ both increase and in this case our algorithm performs no worse than a quadratic time complexity w.r.t. the corpus size. While this could be seen as a potential drawback, it must be contextualised that most privacy preserving algorithms affording strong guarantees including [15,35] often have high time complexities.

Theorem 1. *Given a risklift threshold r_{th}, a set of private terms $S_{\mathcal{P}}$, the serially preserving publication $\mathcal{P}^*$, the anonymised corpus $\mathbb{T}^{\vartheta}_{m+1}$ and our mechanism* Sanony *(π), the serial publication $\mathcal{P}^* + [\pi(\mathbb{T}^{\vartheta}_{m+1})]$ is serially preserving.*

Proof. See Appendix for proof. □

In this section we presented a sound publication mechanism for solving the privacy preserving serial publication problem. We made the observation that composition attacks are only possible if there exists some overlapping transactions between publications. The adversary uses these overlapping transactions to increase the probability of some of the transactions having sensitive terms by excluding others. The new probability of the transactions having a sensitive term upon observing the overlapping transactions is the posterior probability. In order to prevent such knowledge gain without removing the overlapping transactions we introduced a two step approach which makes use of counterfeits. The idea is that in the presence of counterfeits, the chances of excluding some transactions from having a sensitive term is significantly reduced. We demonstrated theoretically that if reasonably small number of counterfeits are added, then the presence of overlapping transactions makes no difference and the posterior probability becomes the same as the prior probability. Our presented approach consists of two main steps, backward perturbation and forward perturbation. Backward perturbation ensures that all transactions of previous publications are at no risk of composition attacks. Forward perturbation has two further parts FP1 which ensures that the current publication cannot be used to conduct any further attacks called transitive composition attacks; and FP2 which ensures that the current publication is serially preserving. We then performed further analysis on our method and showed that it is free from minimality attacks.

In the following, we demonstrate the effectiveness of our proposed method by performing an empirical study.

4. Empirical study

In this section, we demonstrate (1) the susceptibility of transactional data to composition attacks; (2) the perturbation rates of *Sanony*; (3) its utility in count queries; and (4) performance.

4.1. Experimental setup and datasets

Three algorithms were implemented; (1) *Anony*, the non-serial transactional data publication method with a relative risk guarantee as described in Section 2.2 [14]; (2) *Inv*, the serial publication method for relational data with m-invariance guarantee as introduced in Section 1 [7]; and (3) *Sanony*, our proposed solution for serial publication of transactional data with a relative risk guarantee. We also developed *SeGen*, a lightweight serial data generator which samples, with repetition, from a large corpus of transactional data in a serial manner to produce a serial corpora. *SeGen* has tunable parameters such as repeat rate (*rep*) and sample size (*sz*) to allow characterisation of composition attacks within the datasets. All implementations were done in Java and run on an Intel Core i5-4300M CPU @ 2.60 GHz laptop with 8.00 GB RAM and Windows 7 operating system.

Two real world datasets introduced in [36], specifically *BMS-Webview-1*(B1) and *BMS-Webview-2*(B2) were used. Each contains click streams from two online retailers (Table 8). Clicks are represented as terms, and for each dataset we randomly selected 10% of the terms in the corpus as private terms and randomly generated 5 serial corpora $[\mathbb{T}_1, \ldots, \mathbb{T}_5]$ to represent a 5-year serial corpora with *SeGen*. We considered the following parameters:

(1) Sample Size (sz): It is the number of transactions in each serial corpus. *sz* was varied between 1.25% and 10% of the original dataset, with a default of 5%.

(2) Repeat Rate (rep): It is the percentage of transactions of the previous corpus that is also in the subsequent corpus. It was varied between 10% and 80% with a default of 40%.

Table 8
Datasets.

	No. of Trans.	No. of terms	Max Trans. length	Avg. Trans. length	Avg. sparsity
B1	59,602	497	267	2.5	99.49%
B2	77,512	3340	161	5.0	99.86%

(3) Risklift Threshold (r_{th}): It is our privacy parameter (Definition 3). It was varied between 2 (strongest) and 16 (weakest) with a default of 8.

4.2. Composition attacks

This experiment demonstrates the susceptibility of transactional data to composition attacks. For each anonymised corpus in $[\mathbb{T}_1^{\vartheta}, \ldots, \mathbb{T}_5^{\vartheta}]$, the number of transactions at risk was calculated (Definition 12) under varying parameters.

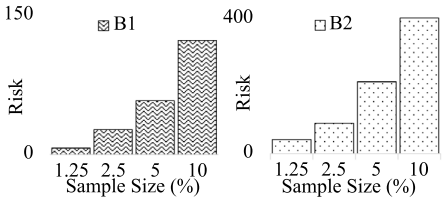
Fig. 6 has 3 pairs of plots (6a, 6b, 6c), each pair has 2 sets of bars corresponding to B1 and B2, and each bar is the total number of transactions at risk for the whole publication. Fig. 6a shows that as the sample size increases, the number of transactions at risk also increases along with it uniformly for both datasets. In general, B2 has more risk compared to B1 because it has more transactions and number of private terms than B1.

Fig. 6b is the results of varying the repeat rate for the serial corpora. As more transactions of the previous corpus $\mathbb{T}_i$ are repeated in the subsequent release $\mathbb{T}_{i+1}$, there is more risk. With more transactions being repeated, there is more chance for overlaps between clusters of different corpora, leading to more inferences. This trend continues until the clusters become identical in both the previous and subsequent corpora and there is no risk. *i.e.* $rep = 100\%$ (not shown in the diagram) there is no risk, as expected.

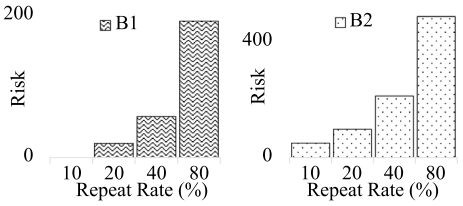
Fig. 6c shows how the number of transactions at risk relates to the privacy parameter r_{th}. Interestingly, with a stricter privacy parameter of $r_{th} = 2$, the number of transactions at risk is lower. This is also true for further smaller values of r_{th} (*i.e.* $r_{th} \in [1, 2]$, not shown in Fig. 6c). As r_{th} becomes bigger, the risk increases until $r_{th} = 8$. This seems counter intuitive but it is explained as follows. At lower (stricter) r_{th}, there is a higher tendency for the anonymisation *Anony* to put more private terms into the global bag in order to satisfy the s-preserving guarantee (Definition 3). In doing so, there are less s terms left in the clusters and fewer overlaps will have any potential to lead to a risk. However as the r_{th} increases (becomes weaker) more s terms remain in their clusters and it is easier for overlaps to result in a higher risk due to the presence of the s terms. This trend continues until r_{th} becomes sufficiently large (around $r_{th} = 16$) such that, even when the risk increases, it no longer breaches the serially preserving criterion *i.e.* $\gamma_s^{ser.}(T^*) \leq r_{th}$ (Definition 12). This characteristic is more pronounced in B1, but is subtle in B2 due to the differences in the datasets (Table 8).

4.3. Perturbation rate

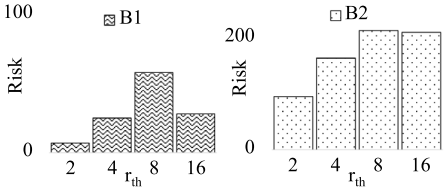
This experiment shows the perturbation rates, *i.e.* the percentage of counterfeits, of *Sanony* in comparison to *Inv* [7]. In *Inv*, the privacy parameter m was set to 2, practically the weakest guarantee of *Inv*, and so lower perturbation rates were expected [7]. In transactions with multiple private terms, the terms were concatenated into one term. The perturbation rate in the resulting *2-invariant* publication was then calculated after applying *Inv* to the serial corpora $[\mathbb{T}_1, \ldots, \mathbb{T}_5]$. Similarly, the perturbation rate of *Sanony* was also calculated. In this comparison we do not involve *Anony* since it does not make use of counterfeits.



(a) Risk vs Sample Size



(b) Risk vs Repeat Rate



(c) Risk vs r_{th}

Fig. 6. Susceptibility to composition attacks.

Fig. 7 which is the results has 3 pairs of plots (7a, 7b, 7c), each pair has 2 sets of bars corresponding to B1 and B2. Each bar represents the perturbation rate of the method, *Sanony* (darker bars) and *Inv* (lighter bars). In Fig. 7a, the perturbation rate w.r.t. sample size is shown. For both datasets B1 and B2, *Sanony* results in much lower perturbation rate. Typically, *Sanony* has a perturbation rate of around 0.2% while *Inv* has 5% for B1. In B2 the average perturbation rate of *Sanony* is 0.5% and it stands at 25% for *Inv*.

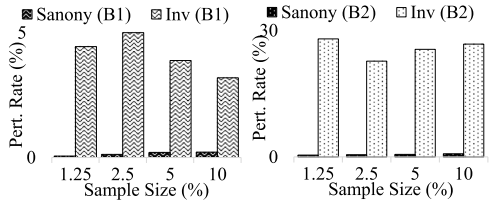
Fig. 7b is the perturbation rate w.r.t. varying repeat rates. For both datasets and methods, the perturbation rates generally increase with the repeat rate due to increasing risk as repeat rate increases (Fig. 6b). However, *Sanony* has significantly lower perturbation rates than *Inv*.

In Fig. 7c, r_{th} was varied in *Sanony* while comparing each result to *Inv* for the fixed privacy parameter of $m = 2$. In the figure, even at its highest perturbation rate ($r_{th} = 8$), *Sanony* performs much better than *Inv*, at least 25 times less perturbation for B1 and 50 times less for B2.

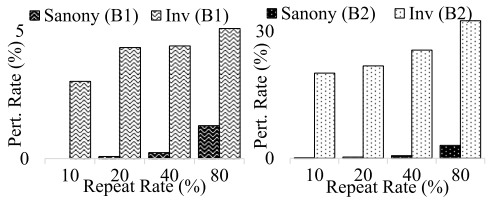
4.4. Utility

This section shows the minimal utility impact of the addition of counterfeits in *Sanony* for actual queries when compared to *Anony*. We note that this is not to compare the utility of *Inv* with *Sanony*, but to show the trade off in using counterfeits to ensure the right privacy guarantee in terms of utility loss. Furthermore, our comparison of perturbation rates in Section 4.3 already demonstrates that *Inv* requires a much higher number of counterfeits. As a result, we compare the error levels of queries for *Anony* and *Sanony* *i.e.* before and after the counterfeits are added respectively.

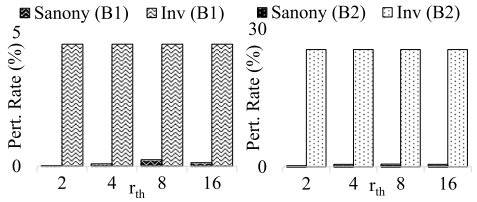
In the experiment, first the serial corpora $[\mathbb{T}_1, \ldots, \mathbb{T}_5]$ was anonymised via *Anony*. Next, for each original corpus $\mathbb{T}_i$, 3 groups



(a) Perturbation Rate vs Sample Size

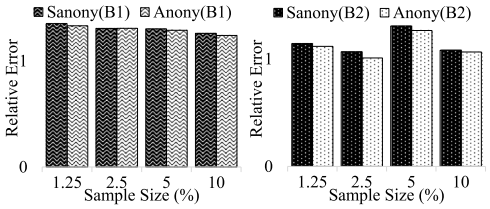


(b) Perturbation Rate vs Repeat Rate

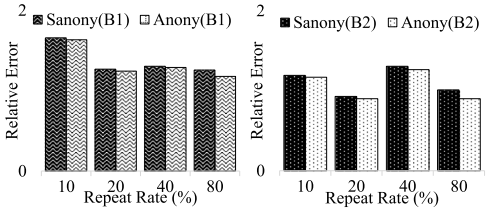


(c) Perturbation Rate vs r_{th}

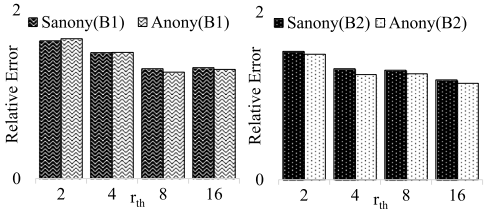
Fig. 7. Perturbation rates.



(a) Utility vs Sample Size



(b) Utility vs Repeat Rate



(c) Utility vs r_{th}

Fig. 8. Utility in queries.

of term associations with respective low (1–10), medium (20–40), and high (70–200) supports in the dataset were randomly selected as queries. The same associations were searched for in their respective anonymised corpus $\mathbb{T}_i^{\vartheta}$ to find their supports. The utility loss of an association (a, b) was measured by the relative error *re* [27]:

$$re = \frac{abs(s_o(a,b) - s_p(a,b))}{AVG(s_o(a,b), s_p(a,b))} \tag{21}$$

where $s_o(a, b)$ and $s_p(a, b)$ are the supports of (a, b) in the original and the anonymised datasets respectively and *abs*() returns the absolute value. The absolute value ensures false associations that did not exist in the original corpus are captured. The denominator uses an average instead of the original support to avoid division by 0 when there are spurious transactions, and it normalises the re values to [0, 2]. We consider only term pairs because we believe they are representative of the more complex term associations that may also be lost by the publication. This is analogous to the *a priori* rule on frequent itemsets [37] *i.e.* if a complex term association is lost, then its term pairs are also lost; and for infrequent complex term associations, the low category of term pairs sufficiently covers them.

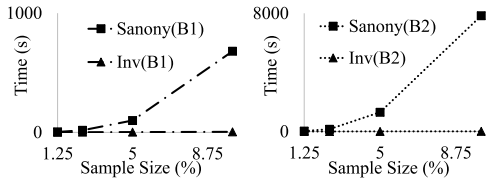
10 different queries were used for each category and the results averaged. For each query, the support from the anonymised data was derived from their reconstructed transactions generated by randomly linking entries in the vertical segments of a cluster (Definition 4). These were generated 20 times and the results averaged. The same experiment was repeated for *Sanony*.

Fig. 8 which is the result has 3 pairs of plots (8a, 8b, 8c), each pair has 2 sets of bars corresponding to B1 and B2. Each bar in a plot is the relative error for a method *Sanony* (darker bars) and *Anony* (lighter bars). In all 3 pairs of plots, we aim to observe the difference between the light and dark bars in each plot. We note that, the actual heights of each compared bar group (light and dark) are not pertinent to our discussion here *i.e.* how the use of counterfeits in *Sanony* to ensure the right privacy guarantee affects utility. The actual heights relate to the cluster formations. For example, different repeat rates may result in different clusters which impacts how *Anony* and *Sanony* perturb the data and subsequently the utility of each bar group [14,38]. The results show only marginal difference in the relative error between *Anony* and *Sanony*. This indicates the impact on utility by *Sanony* is very minimal. In the worst case of Fig. 8b (repeat rate = 80%), the average increase in utility loss is less than 10%. From the results we notice that in some instances ($r_{th} = 2$ of B1 in Fig. 8c) *Sanony* performs slightly better. This is attributed to the randomness in the reconstruction stage as there is the potential that the reconstructed transactions from *Sanony* are closer to the original transactions than *Anony* when $\mathbb{T}_i^{\vartheta}$ and $\pi(\mathbb{T}_i^{\vartheta})$ are similar.

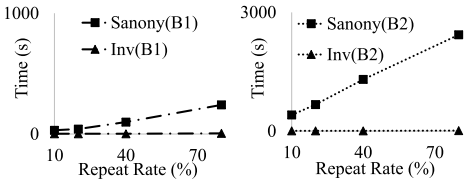
It must be stressed that, this experiment focuses on how much deterioration in utility is caused by *Sanony* when applied on top of *Anony* and may not represent the optimal utility for both methods. It is natural to expect that under optimised utility the same results will hold since the methods employ the same privacy semantics. The interested reader is pointed to [38] for some methods to improve utility in the partitioned based methods.

4.5. Performance of Sanony

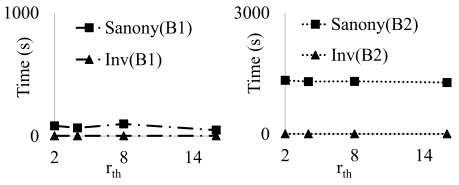
This experiment verifies the performance of *Sanony*. For the same corpora, the runtime of *Sanony* and *Inv* was noted.



(a) Runtime vs Sample Size



(b) Runtime vs Repeat Rate



(c) Runtime vs r_{th}

Fig. 9. Performance.

Fig. 9 which is the results has 3 pairs of plots (9a, 9b, 9c), each plot has 2 lines for *Sanony* and *Inv* respectively. In Fig. 9a, increasing the sample size *sz* for each corpuses in $[\mathbb{T}_1, \ldots, \mathbb{T}_5]$ shows *Sanony* performs no worse than quadratic. As discussed in Section 3.3, this is as a result of the pairwise computation of the overlaps between the corpora. In *Inv*, no such pairwise computation is done and therefore it has a linear time performance. The computation cost associated with *Sanony* is a fair price to pay for its stronger privacy and better perturbation rates. In future, we will investigate the tracking of transactions across publications to reduce the pairwise computations required for determining the overlaps and further establish how our proposal extends to the big data environment as described by [39].

Fig. 9b shows the time behaviour w.r.t. the repeat rate. The results show that for both datasets, the performance is linear w.r.t. the repeat rate for both *Sanony* and *Inv*. This is because, the repeat rate is directly related to the number of transactions at risk (Fig. 7b). Overall, B2 has higher computation cost than B1 because it has more private terms and transactions than B1.

In Fig. 9c, it is seen that r_{th} has no appreciable effect on the performance of *Sanony* for both datasets. Obviously *Inv* is constant since it's privacy parameter is fixed $m = 2$ and this is independent of r_{th}.

5. Related work

Privacy preservation in multiple data publishing has been considered under 3 settings namely *multiple independent publishing*, *collaborative publishing* and *coordinated publishing* [2]. In *multiple independent publishing* each publisher operates independently of other publishers without referencing their datasets. Often, the distribution of private terms in other unknown datasets that may be used for a composition attack is assumed, and used to calculate the risk of any publication [40–42]. In *collaborative publishing* each member of a group of publishers controls a part of the dataset and together wish to jointly publish the overall dataset. Techniques such as secure multiparty computation (SMC) is often used [43–47]. In *coordinated publishing*, a centralised repository where data is updated periodically with new records or attributes is assumed [1, 7–11,48–52].

Our work focuses on *coordinated publishing* and there are two types of publishing schemes [31]; (1) *sequential publishing* considers updates in the form of projections on new attributes; and (2) *serial publishing* considers updates in the form of deletion/insertion of records. In sequential publishing, [48] uses the negative property of *lossy joins* to hide the re-identification of records. [49] extends [48] by considering a local recoding called *cell generalisation* to improve utility. Further [50] improves on [49] by considering a fully dynamic setting.

In serial publishing, [9] and [10] use a delayed publication approach to guarantee k-anonymity for newly inserted records. [1] improves utility by utilising both historical and current data in publishing the current release. [7] introduces persistent invariance in which a record is associated with the same set of private terms, called a signature until it is removed. This effectively preserves m-diversity in the serial publication. In [8], *role composition* provides l-diversity guarantee under a different set of assumptions. Primarily, [8] assumes some values in a record are permanent and do not change while others can be updated arbitrarily. We do not make this assumption as it does not allow updates of some private terms *e.g.* if a person has *HIV* and later develops *T.B.*, his/her record cannot be updated not even with a concatenation $\{HIV + T.B.\}$ *i.e.* it becomes a new private term and defeats the purpose of the permanent private value assumption. In [51], the possibility of privacy risks associated with transient private terms is fully analysed. Further, [11] attempts to solve the problem in [51] by extending m-invariance. [11] assumes, the adversary can track the changes in a transaction via an *event list* to form a τ-attack.

The techniques for serial publishing seen so far adopt relatively weaker privacy guarantees, and they focus on relational data. Recently [52] focussed on spontaneous reporting systems (SRS) under the specific assumption of the presence of *follow-up* keys to link transactions of an individual across publications. Their metric $PPMS(k, \theta^*)$ bounds the probability of any private term s to be linked to an equivalence class of size k to be less than $\theta_s \in \theta$. Unfortunately, its satisfiability cannot be guaranteed *e.g.* given k and θ, there are cases where the probability bound cannot be satisfied. Subsequently, skewness attacks cannot also be prevented when there are rare or combined private terms.

Some elegant utility aware differentially private techniques, which are stronger, also focus mostly on the interactive setting [22,53]. Although [54] presents a formal discussion on differential privacy and composition attacks in the non-interactive setting, how the proposal works in practice is uncertain. The noise required in datasets with rare private values is often prohibitively high [2,28], even for methods such as [55–58] (Fig. 1). Moreover, in [56], *IncTDPart* is proposed to preserve differential privacy in a serial publishing scenario where the dataset updates involves insertions only. [57] also proposes a differentially private counter over binary data streams which does not suitably adapt to the dataset release scenario described in this paper. [58] proposes two techniques *DSFT* and *DSAT*, which produce a differentially private histogram of the current dataset only if it is sufficiently different from the previous dataset to reduce the error introduced by the privacy mechanism. [55] also proposes a hybrid method that makes use of generalisation to reduce the amount of noise added in the case of a mixed relational and set-valued dataset. More recently, [59] considered how temporal correlations between

data values may affect the guarantee of differential privacy, but this solves a different problem.

6. Conclusion

In this work, we considered the privacy preserving serial publication problem of transactional data. When transactional data is published for data analytic applications without care, it can lead to serious privacy disclosures. This problem is further exacerbated in the serial publishing scenario where there are multiple data releases *e.g.* periodic release of prisoner health data [3,4]. This work developed a privacy preserving guarantee for the serial publication of transactional data and then proposed a publication mechanism to satisfy the guarantee.

In developing our privacy guarantee, we first presented the r_{th}-preserving privacy guarantee for the single independent publication scenario [14]. Then, we developed this guarantee further by considering the risk of privacy disclosures (*composition attacks*) based on the posterior knowledge of the adversary. The posterior knowledge of the adversary was defined as the posterior probability of a transaction having a sensitive term after the adversary observes other publications in the serial publication. In calculating this probability, we provided some theoretical reasoning which made use of combinatorics and the Bayes' theorem to maximise the adversary's knowledge. We further analysed the risk by considering the possibility of global compositions, where the adversary can make use the whole of each published corpus to enhance his/her posterior knowledge. Finally, we formally presented the problem this paper sought to address.

We then presented a sound publication mechanism for solving the privacy preserving serial publication problem. We made the observation that composition attacks are only possible if there exists some overlapping transactions between publications. The adversary uses these overlapping transactions to increase the probability of some of the transactions having sensitive terms by excluding others. The new probability of the transactions having a sensitive term upon observing the overlapping transactions is the posterior probability. In order to prevent such knowledge gain without removing the overlapping transactions we introduced a two step approach which makes use of counterfeits. The idea is that in the presence of counterfeits, the chances of excluding some transactions from having a sensitive term is significantly reduced. We demonstrated that if a reasonably small number of counterfeits are added, then the presence of overlapping transactions makes no difference and the posterior probability becomes the same as the prior probability. Our presented approach consists of two main steps, backward perturbation and forward perturbation. Backward perturbation ensures that all transactions of previous publications are at no risk of composition attacks. Forward perturbation has two further parts FP1 which ensures that the current publication cannot be used to conduct other attacks such as the transitive composition attacks; and FP2 which ensures that the current publication is serially preserving. We then performed further analysis on our method and showed that it is free from minimality attacks.

Finally, we demonstrated the effectiveness of our method on real datasets and in comparison with state-of-the-art methods. In particular, it was seen that single independent publication mechanisms are prone to composition attacks and our method effectively defends against such attacks. We also demonstrated that our method gave significantly better results in terms of lower perturbation rates than other methods such as m-invariance [7] which also makes use counterfeits. In addition, when we compared our method to the single independent publication method *Anony* that does not use counterfeits, we saw that the effect of the counterfeits added in our method *Sanony* was marginal (less than 10% increase in utility loss in the worst case). These experiments were based on the error rate of actual queries.

Acknowledgement

This research work has been partially supported by ARC, Australia Discovery grant DP140103617.

Appendix

A.1. Proof of Lemma 1

Proof. In Formula (12) it is clear that each overlap $\mathbf{O}_i \in \Omega(\mathbf{C}^*)$ derives an inference about the private term s of T^*; and the effect of all the overlaps must be considered to avoid a bias in calculating the true posterior probability of T^*. □

A.2. Proof of Lemma 2

Proof. This lemma holds because of Formulae (13) and (14). In the formulae, $Prob(\mathbf{O}|s \in T^r) \longrightarrow 1\ (1^\sim)$ and $Prob(\mathbf{O}|s \notin T^r) \longrightarrow 1\ (1^\sim)$ as the interval of $[r_1, r_2]_s$ increases. Formula (12) becomes $Prob(s \in T^r|\mathbf{O}) = \frac{Prob(s\in T^r)\cdot 1^\sim}{1^\sim \cdot Prob(s\in T^r)+1^\sim \cdot Prob(s\notin T^r)}$. Clearly, $Prob(s \in T^r|\mathbf{O}) \longrightarrow Prob(s \in T^r)$ since the denominator $(1^\sim \cdot Prob(s \in T^r) + 1^\sim \cdot Prob(s \notin T^r)) \longrightarrow 1$. □

While the overlap relates to both $\mathbf{C}^*$ and $\mathbf{C}^\vartheta$, Lemma 2 applies only to $\mathbf{C}^*$ because the proof is true for fixed cluster *i.e.* for $T^* \in \mathbf{C}^*$ if $\beta(T^*, s) = \alpha(T^*, s)$ is true due to Lemma 2 and an s-transaction is added to $\mathbf{C}^*$ then $\beta(T^*, s) = \alpha(T^*, s)$ may NO longer be true. However, $\mathbf{C}^*$ is already published and no new transactions can be added to it, so Lemma 2 is correct.

A.3. Proof of Lemma 3

Proof. In Formula (13) *Vandermonde's identity* [60] implies that:

$$\sum_{r=0}^{|\mathbf{O}-z|} \binom{N(s,\mathbf{C}^*)-1}{r-z} \cdot \binom{N(\bar{s},\mathbf{C}^*)}{N(\mathbf{O})-r} = \binom{N(\mathbf{C}^*)-1}{N(\mathbf{O})-z}$$

since $(r-z)+(N(\mathbf{O})-r) = N(\mathbf{O})-z$ and $(N(s,\mathbf{C}^*)-1)+N(\bar{s},\mathbf{C}^*) = N(\mathbf{C}^*)-1$ for all their non-negative values. The *identity* is also true for all values $r \in [f_{1_{\mathbf{C}^*}}, f_{2_{\mathbf{C}^*}}]_s$ *i.e.*

$$\sum_{r=f_{1_{\mathbf{C}^*}}}^{f_{2_{\mathbf{C}^*}}} \binom{N(s,\mathbf{C}^*)-1}{r-z} \cdot \binom{N(\bar{s},\mathbf{C}^*)}{N(\mathbf{O})-r} = \binom{N(\mathbf{C}^*)-1}{N(\mathbf{O})-z}$$

For any value $r \notin [f_{1_{\mathbf{C}^*}}, f_{2_{\mathbf{C}^*}}]_s$ an impossible combination of $\binom{N(\bar{s},\mathbf{C}^*)}{N(\mathbf{O})-r}$ or $\binom{N(s,\mathbf{C}^*)-1}{r-z}$ results which evaluates to 0 so the *identity* is trivially satisfied. The probability $Prob(\mathbf{O}|s \in T^r)$ in Formula (13) then becomes 1. Similarly it can also be shown that $Prob(\mathbf{O}|s \notin T^r)$ in Formula (14) also becomes 1; and $Prob(s \in T^r|\mathbf{O}) = Prob(s \in T^r)$ in Formula (12). □

A.4. Proof of Lemma 4

Proof. We need to show that, the addition of $h_{\bar{s}}$ $\bar{s}$-counterfeits and h_s s-counterfeits to $\mathbf{C}^\vartheta$ transforms the range $[r_1, r_2]_s$ to $[r'_1, r'_2]_s$ such that $f_{1_{\mathbf{C}^*}} \geq r'_1$ and $f_{2_{\mathbf{C}^*}} \leq r'_2$ to satisfy Lemma 3. Let the s-range of $\mathbf{O}$ w.r.t. $\mathbf{C}^\vartheta$ be $[f_{1_{\mathbf{C}^\vartheta}}, f_{2_{\mathbf{C}^\vartheta}}]_s$. $h_{\bar{s}} > 0$ implies $r_1 = f_{1_{\mathbf{C}^\vartheta}} > f_{1_{\mathbf{C}^*}}$ (Formula (6)). In Formula (5), $f_{1_{\mathbf{C}^\vartheta}}$ reduces as $\bar{s}$-counterfeits are added since $N(\mathbf{C}^\vartheta \setminus \mathbf{O}^{\bar{S}})$ increases but $N(\mathbf{O})$ never increases due to Definition 14. So adding $h_{\bar{s}} = r_1 - f_{1_{\mathbf{C}^*}}$ $\bar{s}$-counterfeits to $\mathbf{C}^\vartheta$ makes $f_{1_{\mathbf{C}^\vartheta}} \leq f_{1_{\mathbf{C}^*}}$

true and $r'_1 = f_{1_{\mathbf{C}^*}}$. When $h_s > 0$ it implies $r_2 = f_{2_{\mathbf{C}^\vartheta}} < f_{2_{\mathbf{C}^*}}$ (Formula (6)). As $f_{2_{\mathbf{C}^\vartheta}}$ depends solely on $N(s, \mathbf{C}^\vartheta)$ (Formula (5)), adding $h_s = f_{2_{\mathbf{C}^*}} - r_2$ s-counterfeit to $N(s, \mathbf{C}^\vartheta)$ makes $f_{2_{\mathbf{C}^\vartheta}} = f_{2_{\mathbf{C}^*}}$ and $r'_2 = f_{2_{\mathbf{C}^*}} = f_{2_{\mathbf{C}^\vartheta}}$. Lemma 3 is satisfied so Lemma 4 is correct. □

A.5. Proof of Lemma 5

Proof. Let $[r_1, r_2]_s$ be the range of s in $\mathbf{O}$ and $[f_{1_{\mathbf{C}^\mu}} = f_1(\mathbf{C}^\mu, \mathbf{d}_o^{\bar{s}}), f_{2_{\mathbf{C}^\mu}} = f_2(\mathbf{C}^\mu, \mathbf{d}_o^{\bar{s}})]$ be the s-range of the true non-overlapping transactions $\mathbf{d}_o$ after FP1 (Formula (5)). Also let $x_c = d_s + d_{\bar{s}}$ be the number of counterfeits added in FP1. We need to show that the addition of x_c counterfeits, transforms $[r_{1_d}, r_{2_d}]_s$ to $[r'_{1_d}, r'_{2_d}]_s$ so that $f_{1_{\mathbf{C}^\mu}} \geq r'_{1_d}$ and $f_{2_{\mathbf{C}^\mu}} \leq r'_{2_d}$ and there can be no inference (Lemma 3). We begin with the following formulae derived from Formulae (19) and (5) respectively.

$$r'_{1_d} = \max\{r_{1_f} - (x_d + x_c), 0\}$$
$$r'_{2_d} = \min\{(N(\mathbf{C}^\mu) - (x_d + x_c) - r_1), r_{2_f}\}$$
$$f_{1_{\mathbf{C}^\mu}} = \max\Big\{\big(N(s, \mathbf{C}^\mu) - N(\mathbf{C}^{\bar{S}} \setminus \mathbf{d}_o^{\bar{s}})\big), 0\Big\}$$
$$f_{2_{\mathbf{C}^\mu}} = \min\Big\{N(\mathbf{d}_o), N(s, \mathbf{C}^\mu)\Big\}$$

The addition of x_c makes $r'_{1_d} = 0$. In r'_{1_d}, $r_{1_f} = N(s, \mathbf{C}^\mu) - r_2$ (Formula (8)) and $x_c = (N(s, \mathbf{C}^\varrho) - r_2 - x_d) + d_s$ so $r'_{1_d} = N(s, \mathbf{C}^\mu) - r_2 - (x_d + N(s, \mathbf{C}^\varrho) - r_2 - x_d + d_s) = 0$ as $N(s, \mathbf{C}^\mu) = N(s, \mathbf{C}^\varrho) + d_s$. So $f_{1_{\mathbf{C}^\mu}} = r'_{1_d}$ is true. The addition of x_c makes $f_{2_{\mathbf{C}^\mu}} = N(\mathbf{d}_o)$. In $f_{2_{\mathbf{C}^\mu}}$, $N(s, \mathbf{C}^\mu) = N(s, \mathbf{C}^\varrho) + d_s$ and $d_s = N(\mathbf{d}_o) + r_1 - N(s, \mathbf{C}^\varrho)$ (Formula (20)) so $N(s, \mathbf{C}^\mu) = N(\mathbf{d}_o) + r_1$ and $f_{2_{\mathbf{C}^\mu}} = \min\{N(\mathbf{d}_o), N(\mathbf{d}_o) + r_1\} = N(\mathbf{d}_o)$. In r'_{2_d}, $r_{2_f} = N(\mathbf{d}_o) + r_1 - r_1 = N(\mathbf{d}_o)$ as $N(s, \mathbf{C}^\mu) = N(\mathbf{d}_o) + r_1$. Also $(N(\mathbf{C}^\mu) - (x_d + x_c) - r_1) = N(\mathbf{C}^\mu) - r_1$ and $r'_{2_d} = \min\{(N(\mathbf{C}^\mu) - r_1), N(\mathbf{d}_o)\} = \mathbf{d}_o^{\bar{S}}$ since $N(\mathbf{C}^\mu) - r_1 \geq N(\mathbf{d}_o)$ ($\mathbf{d}_o^{\bar{S}} = \mathbf{C}^{\bar{S}} \setminus \mathbf{O}$ ($\mathbf{C}^{\bar{S}} \in \mathbf{C}^\vartheta$) & $r_1 \leq N(\mathbf{O})$), so $f_{2_{\mathbf{C}^\mu}} = r'_{2_d}$ and Lemma 5 is proved. □

From the proof, it is seen that the addition of more counterfeits (both s and $\bar{s}$) does not affect $f_{1_{\mathbf{C}^\mu}} = r'_{1_d}$ or $f_{2_{\mathbf{C}^\mu}} = r'_{2_d}$, Lemma 3 remains satisfied and Lemma 5 is correct.

A.6. Proof of Lemma 6

Proof. We focus on the transactions in the derived clusters since those in the overlap are serially preserving due to Lemma 3. We need to show that the addition of $\bar{s}$-counterfeits causes the posterior probability to decrease up to a point where T^* becomes serially preserving; and the number of counterfeits added x_c is maximally θ_c.

We simplify the notation for this proof as follows: $P\bar{O} = Prob(\mathbf{O}|s \notin T^r)$ and $PO = Prob(\mathbf{O}|s \in T^r)$. $ns = N(s, \mathbf{C}^*)$ and $n\bar{s} = N(\bar{s}, \mathbf{C}^*)$. Also, $Prob(s \in T^*|\mathbf{O}) = 1/(1 + \frac{P\bar{O} \cdot Prob(s \notin T^r)}{PO \cdot Prob(s \in T^r)})$ (Formula (12)), we focus on the ratio $\frac{P\bar{O} \cdot Prob(s \notin T^r)}{PO \cdot Prob(s \in T^r)} = \frac{\frac{n\bar{s}}{N(\mathbf{C}^*)} \cdot \sum_{r=r_1}^{r_1} \binom{ns}{r} \cdot \binom{n\bar{s}-1}{|\mathbf{O}|-r-z}}{\frac{ns}{N(\mathbf{C}^*)} \cdot \sum_{r=r_1}^{r_1} \binom{ns-1}{r-z} \cdot \binom{n\bar{s}}{|\mathbf{O}|-r}}$ (Formula (13) and (14)). For $r \in [r_1, r_2]_s$ and $z = 0$ (since we focus on the derived cluster and $T^* \notin \mathbf{O}$), the ratio simplifies to $\frac{n\bar{s}-o+r}{ns-r}$ by *factorial expansion*. Let $\frac{P\bar{O}' \cdot Prob(s \notin T^r)'}{PO' \cdot Prob(s \in T^r)'} = \frac{\frac{n\bar{s}+1}{N(\mathbf{C}^*)+1} \cdot \binom{ns}{r} \cdot \binom{(n\bar{s}+1)-1}{|\mathbf{O}|-r-z}}{\frac{ns}{N(\mathbf{C}^*)+1} \cdot \binom{ns-1}{r-z} \cdot \binom{n\bar{s}+1}{|\mathbf{O}|-r}}$ be the new ratio after an $\bar{s}$-counterfeit is added. It also simplifies to $\frac{n\bar{s}-\mathbf{O}+r+1}{ns-r}$ which is greater than $\frac{n\bar{s}-o+r}{ns-r}$. Therefore, the posterior probability reduces since $\frac{P\bar{O}' \cdot Prob(s \notin T^r)'}{PO' \cdot Prob(s \in T^r)'} > \frac{P\bar{O} \cdot Prob(s \notin T^r)}{PO \cdot Prob(s \in T^r)}$.

The number x_c of $\bar{s}$-counterfeits required can be shown to be maximally equal to θ_c, by considering the worst case scenario where the overlap $\mathbf{O}$ and the cluster $\mathbf{C}^\mu$ differ on one transaction *i.e.* $N(\mathbf{C}^\mu) = N(\mathbf{O}) + 1$ and it is associated with the minimum copies of s, *i.e.* $r = r_2 = r_1$ so the derived cluster has the maximum copies of s (Proposition 2). If the posterior probability is A/B, reducing A/B to satisfy the serially preserving condition $A/B \leq ss(s, \mathbb{T}^*) \cdot r_{th}$ requires $\theta_c = \frac{A}{ss(s, \mathbb{T}^\mu) \cdot r_{th}} - B$ by *algebraic* manipulation. □

It is worth mentioning that, as expected, when $T^* \in \mathbf{O}$, *i.e.* $z = 1$, then $\frac{P\bar{O} \cdot Prob(s \notin T^r)}{PO \cdot Prob(s \in T^r)} = \frac{P\bar{O}' \cdot Prob(s \notin T^r)'}{PO' \cdot Prob(s \in T^r)'}$ is always true in the above proof (for both s and $\bar{s}$ counterfeits). This confirms that the transactions in the overlap are always serially preserving due to Lemma 3.

A.7. Proof of Theorem 1

Proof. After BP, each transaction in the previous publication is serially preserving due to Lemma 3. After FP1, there can be no transitive attacks because of Lemma 5. And due to Lemma 6, FP2 ensures that all the transactions of $\pi(\mathbb{T}^\vartheta_{m+1})$ are serially preserving, and the addition of further counterfeits does not affect Lemmas 3 and 5. Theorem 1 is correct. □

References

[1] B.C.M. Fung, K. Wang, A.W. Fu, J. Pei, Anonymity for continuous data publishing, in: EDBT 2008, 11th International Conference on Extending Database Technology, Nantes, France, March 25-29, 2008. Proceedings, 2008, pp. 264–275.

[2] J. Li, S.A. Sattar, M.M. Baig, J. Liu, R. Heatherly, Q. Tang, B. Malin, Methods to mitigate risk of composition attack in independent data publications, in: Medical Data Privacy Handbook, 2015, pp. 179–200.

[3] T. Butler, L. Boonwaat, S. Hailstone, National Prison Entrants' Bloodborne Virus Survey, Centre for Health Research in Criminal Justice & National Centre in HIV Epidemiology and Clinical Research, University of New South Wales, 2005.

[4] Anon. H.I.V., Viral Hepatitis and Sexually Transmissible Infections in Australia Annual Surveillance Report, The Kirby Institute, The University of New South Wales, 2013.

[5] X. Xiao, Y. Tao, Anatomy: simple and effective privacy preservation, in: Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006, 2006, pp. 139–150.

[6] N. Li, T. Li, S. Venkatasubramanian, Closeness: a new privacy measure for data publishing, IEEE Trans. Knowl. Data Eng. 22 (7) (2010) 943–956.

[7] X. Xiao, Y. Tao, M-invariance: towards privacy preserving re-publication of dynamic datasets, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China, June 12-14, 2007, 2007, pp. 689–700.

[8] Y. Bu, A.W. Fu, R.C. Wong, L. Chen, J. Li, Privacy preserving serial data publishing by role composition, in: PVLDB '08, Proceedings of 34th International Conference on Very Large Data Bases, August 23-28, 2008, Vol. 1, Auckland, New Zealand, 2008, pp. 845–856, (1).

[9] J. Byun, Y. Sohn, E. Bertino, N. Li, Secure anonymization for incremental datasets, in: Secure Data Management, Third VLDB Workshop, SDM 2006, Seoul, Korea, September 10-11, 2006. Proceedings, 2006, pp. 48–63.

[10] J. Pei, J. Xu, Z. Wang, W. Wang, K. Wang, Maintaining k-anonymity against incremental updates, in: 19th International Conference on Scientific and Statistical Database Management, SSDBM 2007, 9-11 2007, Banff, Canada, Proceedings, 2007, p. 5.

[11] A. Anjum, G. Raschia, M. Gelgon, A. Khan, S.U.R. Malik, N. Ahmad, M. Ahmed, S. Suhail, M. Alam, τ-safety: a privacy model for sequential publication with arbitrary updates, Comput. Secur. 66 (2017) 20–39.

[12] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, L-diversity: privacy beyond k-anonymity, in: Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 2006, Atlanta, GA, USA, 2006, p. 24.

[13] L. Sweeney, K-anonymity: a model for protecting privacy, Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 10 (5) (2002) 557–570.

[14] M. Bewong, J. Liu, L. Liu, J. Li, K.R. Choo, A relative privacy model for effective privacy preservation in transactional data, in: 2017 IEEE Trustcom/BigDataSE/ICESS, Sydney, Australia, August 1-4, 2017, 2017, pp. 394–401.

[15] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, S. Martínez, T-closeness through microaggregation: strict privacy with enhanced utility preservation, IEEE Trans. Knowl. Data Eng. 27 (11) (2015) 3098–3110.

[16] C. Dwork, Differential privacy, in: ICALP, 2006, pp. 1–12.

[17] D. Kifer, A. Machanavajjhala, No free lunch in data privacy, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, Athens, Greece, June 12-16, 2011, 2011, pp. 193–204.

[18] G. Cormode, Personal privacy vs population privacy: learning to attack anonymization, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011, 2011, pp. 1253–1261.

[19] D. Kifer, A. Machanavajjhala, Pufferfish: a framework for mathematical privacy definitions, ACM Trans. Database Syst. 39 (1) (2014) 3:1–3:36.
[20] S. Haney, A. Machanavajjhala, J.M. Abowd, M. Graham, M. Kutzbach, L. Vilhuber, Utility cost of formal privacy for releasing national employer-employee statistics, in: Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017, 2017, pp. 1339–1354.
[21] B. Yang, I. Sato, H. Nakagawa, Bayesian differential privacy on correlated data, in: ACM SIGMOD, 2015, pp. 747–762.
[22] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, D. Megías, Individual differential privacy: a utility-preserving formulation of differential privacy guarantees, IEEE Trans. Inf. Forensics Secur. 12 (6) (2017) 1418–1429.
[23] C. Dwork, Differential privacy: a survey of results, in: TAMC, 2008, pp. 1–19.
[24] C. Dwork, Ask a better question, get a better answer a new approach to private data analysis, in: Database Theory - ICDT 2007, 11th International Conference, Barcelona, Spain, January 10-12, 2007. Proceedings, 2007, pp. 18–27.
[25] R. Chen, N. Mohammed, B.C.M. Fung, B.C. Desai, L. Xiong, Publishing set-valued data via differential privacy, PVLDB 4 (11) (2011) 1087–1098.
[26] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, S. Martínez, Enhancing data utility in differential privacy via microaggregation-based $$k$$ k -anonymity, VLDB J. 23 (5) (2014) 771–794.
[27] M. Terrovitis, J. Liagouris, N. Mamoulis, S. Skiadopoulos, Privacy preservation by disassociation, in: PVLDB 12, Proceedings of 38th International Conference on Very Large Data Bases, August 27-31, 2012. Instanbul, Turkey, Vol. 5, 2012, pp. 944–955, (10).
[28] M. Bewong, J. Liu, L. Liu, J. Li, K.R. Choo, A relative privacy model for effective privacy preservation in transactional data, Concurr. Comput.: Pract. Exper. 9 (2018).
[29] J. Soria-Comas, J. Domingo-Ferrer, Differential privacy via t-closeness in data publishing, in: Eleventh Annual International Conference on Privacy, Security and Trust, PST 2013, 10-12 July, 2013, Tarragona, Catalonia, Spain, July 10-12, 2013, 2013, pp. 27–35.
[30] Anon, Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (hipaa) Privacy Rule, U.S. Department of Health and Human Services, 2012.
[31] B.C.M. Fung, K. Wang, R. Chen, P.S. Yu, Privacy-preserving data publishing: a survey of recent developments, ACM Comput. Surv. 42 (4) (2010).
[32] R.C. Wong, A.W. Fu, K. Wang, J. Pei, Minimality attack in privacy preserving data publishing, in: Proceedings of the 33rd International Conference on Very Large Data Bases, University of Vienna, Austria, September 23-27, 2007, 2007, pp. 543–554.
[33] X. Xiao, Y. Tao, N. Koudas, Transparent anonymization: thwarting adversaries who know the algorithm, ACM Trans. Database Syst. 35 (2) (2010) 8:1–8:48.
[34] S.J. Russell, P. Norvig, Artificial Intelligence - A Modern Approach, third internat. ed., Pearson Education, 2010.
[35] G. Loukides, A. Gkoulalas-Divanis, B. Malin, cOAT: constraint-based anonymization of transactions, Knowl. Inf. Syst. 28 (2) (2011) 251–282.
[36] Z. Zheng, R. Kohavi, L. Mason, Real world performance of association rule algorithms, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 26-29, 2001, 2001, pp. 401–406.
[37] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile, 1994, pp. 487–499.
[38] M. Bewong, J. Liu, L. Liu, J. Li, Utility aware clustering for publishing transactional data, in: Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017. Proceedings, Part II, 2017, pp. 481–494.
[39] H. Zakerzadeh, C.C. Aggarwal, K. Barker, Privacy-preserving big data publishing, in: Proceedings of the 27th International Conference on Scientific and Statistical Database Management, SSDBM '15, La Jolla, CA, USA, June 29 - July 1, 2015, 2015, pp. 26:1–26:11.
[40] M.M. Baig, J. Li, J. Liu, H. Wang, Cloning for privacy protection in multiple independent data publications, in: Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011, 2011, pp. 885–894.
[41] A.H.M.S. Sattar, J. Li, J. Liu, R. Heatherly, B. Malin, A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments, Knowl.-Based Syst. 67 (2014) 361–372.
[42] J. Li, M.M. Baig, A.H.M.S. Sattar, X. Ding, J. Liu, M.W. Vincent, A hybrid approach to prevent composition attacks for independent data releases, Inform. Sci. 367–368 (2016) 324–336.
[43] W. Jiang, C. Clifton, A secure distributed framework for achieving k-anonymity, VLDB J. 15 (4) (2006) 316–333.
[44] D. Alhadidi, N. Mohammed, B.C.M. Fung, M. Debbabi, Secure distributed framework for achieving ϵ-differential privacy, in: Privacy Enhancing Technologies - 12th International Symposium, PETS 2012, Vigo, Spain, July 11-13, 2012. Proceedings, 2012, pp. 120–139.
[45] N. Mohammed, B.C.M. Fung, K. Wang, P.C.K. Hung, Privacy preserving data mashup, in: EDBT 2009, 12th International Conference on Extending Database Technology, Saint Petersburg, Russia, March 24-26, 2009. Proceedings, 2009, pp. 228–239.
[46] S. Su, P. Tang, X. Cheng, R. Chen, Z. Wu, Differentially private multi-party high-dimensional data publishing, in: 32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016, 2016, pp. 205–216.
[47] J. Hua, A. Tang, Y. Fang, Z. Shen, S. Zhong, Privacy-preserving utility verification of the data published by non-interactive differentially private mechanisms, IEEE Trans. Inf. Forensics Secur. 11 (10) (2016) 2298–2311.
[48] K. Wang, B.C.M. Fung, Anonymizing sequential releases, in: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006, 2006, pp. 414–423.
[49] E. Shmueli, T. Tassa, R. Wasserstein, B. Shapira, L. Rokach, Limiting disclosure of sensitive data in sequential releases of databases, Inform. Sci. 191 (2012) 98–127.
[50] E. Shmueli, T. Tassa, Privacy by diversity in sequential releases of databases, Inform. Sci. 298 (2015) 344–372.
[51] R.C. Wong, A.W. Fu, J. Liu, K. Wang, Y. Xu, Global privacy guarantee in serial data publishing, in: Proceedings of the 26th International Conference on Data Engineering, ICDE 2010, March 1-6, 2010, Long Beach, California, USA, 2010, pp. 956–959.
[52] J. Wang, W. Lin, Privacy preserving anonymity for periodical SRS data publishing, in: 33rd IEEE International Conference on Data Engineering, ICDE 2017, San Diego, CA, USA, April 19-22, 2017, 2017, pp. 1344–1355.
[53] H. Li, L. Xiong, Z. Ji, X. Jiang, Partitioning-based mechanisms under personalized differential privacy, in: Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017. Proceedings, Part I, 2017, pp. 615–627.
[54] S.R. Ganta, S.P. Kasiviswanathan, A.D. Smith, Composition attacks and auxiliary information in data privacy, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008, 2008, pp. 265–273.
[55] N. Mohammed, X. Jiang, R. Chen, B.C.M. Fung, L. Ohno-Machado, Privacy-preserving heterogeneous health data sharing, JAMIA 20 (3) (2013) 462–469.
[56] X. Zhang, X. Meng, R. Chen, Differentially private set-valued data release against incremental updates, in: Database Systems for Advanced Applications, 18th International Conference, DASFAA 2013, Wuhan, China, April (2013) 22-25. Proceedings, Part I, 2013, pp. 392–406.
[57] C. Dwork, M. Naor, T. Pitassi, G.N. Rothblum, Differential privacy under continual observation, in: Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 2010, 2010, pp. 715–724.
[58] H. Li, L. Xiong, X. Jiang, J. Liu, Differentially private histogram publication for dynamic datasets: an adaptive sampling approach, in: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19-23, 2015, 2015, pp. 1001–1010.
[59] Y. Cao, M. Yoshikawa, Y. Xiao, L. Xiong, Quantifying differential privacy under temporal correlations, in: 33rd IEEE International Conference on Data Engineering, ICDE 2017, San Diego, CA, USA, April 19-22, 2017, 2017, pp. 821–832.
[60] H.W. Gould, A new symmetrical combinatorial identity, Combin. Theory A (13) (1972) 278–286.