PRIVACY PRSERVING ATTRIBUTION AND MEASUREMENT

IMDA PET SANDBOX – TIKTOK CASE STUDY

INFOCOMM
MEDIA
DEVELOPMENT
AUTHORITY

Contents

Use Case Background

1. To ensure that consumers are served relevant advertisements, data on customer behaviour and activity is essential. **Third party cookies have traditionally fulfilled this information requirement by enabling advertisers to track and link the behaviour of users across websites**, allowing them to understand the profile of customers to whom they serve advertisements. However, growing global privacy concerns could lead to calls for the eventual deprecation of 3rd party cookies.

2. In the absence of 3rd party cookies, there would still be a need to understand consumer activity and behaviour**. Publishers and advertisers would need to find a mechanism or tool that enables them to learn more about their consumers but in a manner that preserves the privacy of the consumer**, i.e. not being able to re-identify or discern who the individual may be.

3. **TikTok has proposed "PrivacyGo",** an open-sourced solution that leverages the use of various PETs, to enable publishers and advertisers to compute aggregate measurements of advertising campaigns efficiently and accurately while ensuring the privacy of individual users.

Data Sharing Considerations

4. Calculation of advertisement attribution & measurement requires consumer data from both publishers and advertisers, but this is not data that can readily be shared between companies as customers' information is involved.

Use Case Details

5. A Proof of Concept (POC) was conducted seeking to evaluate the impact of PETs on user privacy and data security across the process of advertisement attribution

6. POC stakeholders included:
   a. **TikTok (publisher) -** holds identifiers corresponding to its users who have viewed the advertising campaign.
   b. **Mobility service provider (advertiser)** - holds identifiers and numeric values relating to their users who have made purchases and corresponding spending amounts.

POC Design Considerations

a. **Size of dataset** - The computation should be able to support large-scale customer databases. (e.g. 10 million.)

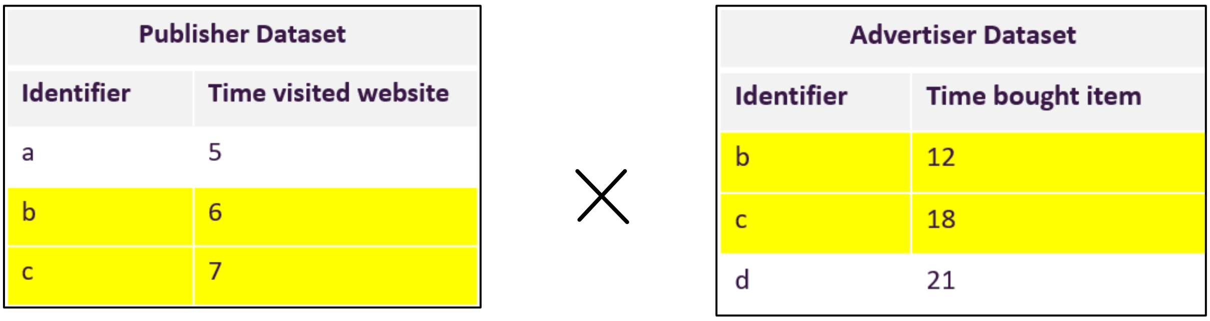
b. **Compute time** – The computation should be completed within a reasonable amount of time (e.g. in hours) using commercially available servers.

c. **Communication cost (i.e. network bandwidth traffic)** – Solution should be designed such that the cost to transfer data between nodes should also be as small as possible

Solution Overview

7. The solution consists of 2 key components: Intersect and Compute.

a. **Intersect:** Match common customers from private data between the publisher and the advertiser. This leverages DPCA-PSI[1] protocol, which is a specifically designed protocol combining ECDH-style PSI and a differentially private mechanism. It not only enables secure multi-ID matching but also obfuscates intersection sizes, thereby mitigating risks including membership inference attacks.



Publisher Dataset	
Identifier	**Time visited website**
a	5
b	6
c	7

×

Advertiser Dataset	
Identifier	**Time bought item**
b	12
c	18
d	21

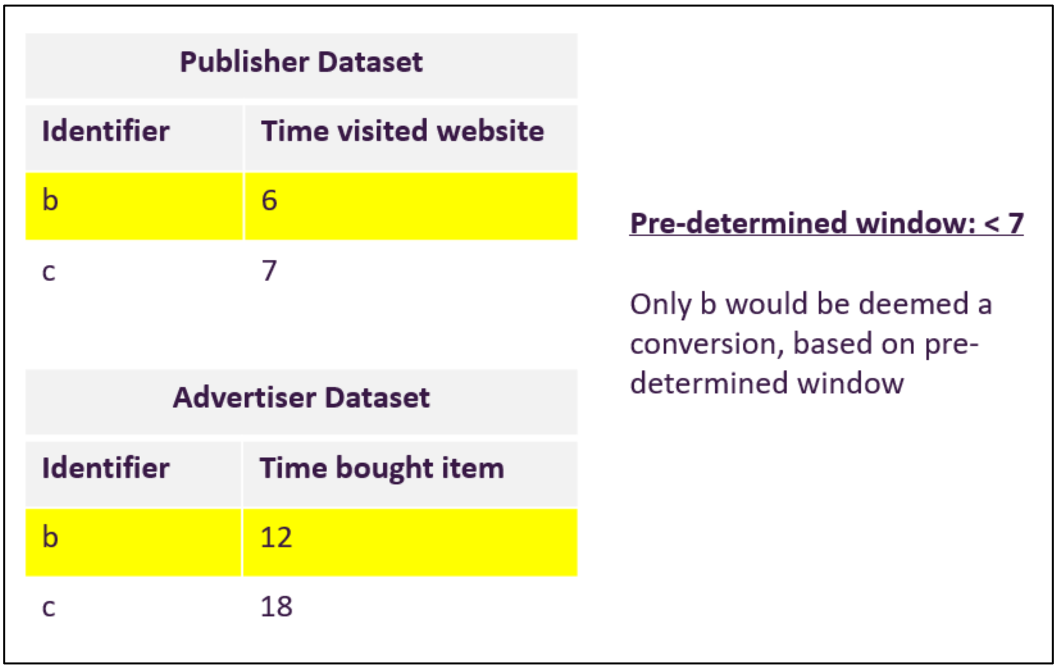
Intersecting between publisher and advertiser datasets

b. **Compute:** Calculate the qualified conversion values for valid transactions. Valid transaction is defined as conversions that happened within a pre-determined window after impressions (i.e. only if a consumer purchases a product within X no. of seconds

[1] ECDH-style PSI is widely used in the industry due to their lower communication complexity, especially for large-scale data handling and multi-ID matching.

[2] MPC-DualDP is a distributed protocol for generating shared differential privacy noise in a two-server setting. MPC-DualDP leverages MPC to sample random noise according to specific distributions, and outputs the noise in the form of secret sharing.

after seeing the advertisement) while protecting the conversion values and learning output of the computation. This component leverages MPC-DualDP to protect the individual conversion values and its output. In particular, Differential Privacy is added such that MPC outputs are perturbed by the addition of noise before outputs are revealed.



Publisher Dataset	
Identifier	**Time visited website**
b	6
c	7

Advertiser Dataset	
Identifier	**Time bought item**
b	12
c	18

Pre-determined window: < 7

Only b would be deemed a conversion, based on pre-determined window

Computing conversion between publisher and advertiser datasets

POC Steps

8. There are 2 phases to the POC:

Phase 1 – Matching of common customers

a. **Step 1** – Before identifying the datasets that the publisher and advertiser would be using to conduct attribution analysis, two data protection measures were implemented to reduce the risk of re-identification:
    a. Addition of randomly generated noise to the datasets to create false matches
    b. Rows within datasets are shuffled, to garble the original orientation

b. **Step 2a** - Identifiers within both publisher and advertiser datasets are then double encrypted using elliptic curve cryptography (ECC) (i.e. customer IDs are encrypted with both parties’ private keys). Double encryption enabled Private Set Intersection (PSI), to identify common customers in both datasets without revealing data or exposing sensitive information (i.e. who exact common customers are) to either party.

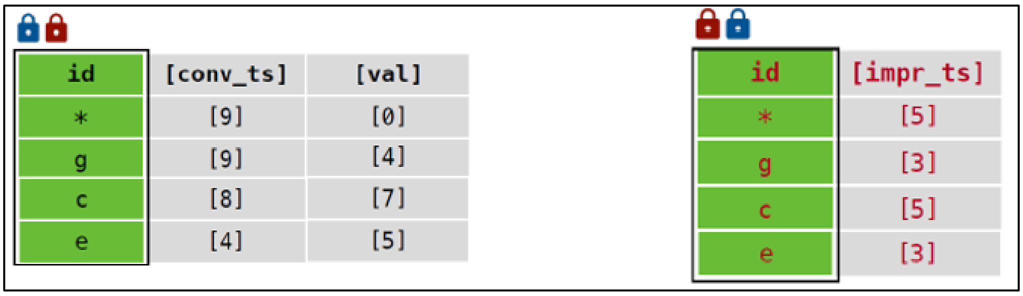
c. **Steps 2b** – Accompanying attributes (e.g. impression timestamp, conversion timestamp and conversion value) are also homomorphically encrypted (HE) to enable the creation of encrypted secret shares in step 4. The use of HE also serves as an additional data protection safeguard to ensure attributes cannot be seen by another party, but can still be processed

<u>Homomorphic Encryption</u>

a. TikTok and Advertiser both additionally "hide" their customers' data using homomorphic encryption ("HE") through the Paillier encryption[3] algorithm (defined as an encryption technique in the ISO/IEC standard under 18033.6), prior to the generation of secret shares.

b. HE allows data to be computed in an encrypted state, thus ensuring data is not revealed to anyone even when in use. HE is used to prepare for the next secret sharing stage by creating encrypted secret shares of matched customers' conversion values through SMPC. These encrypted secret shares are then returned to and decrypted by the originating party to reveal the "plaintext" secret share conversion value.

d. **Steps 3** – Common customers are found through matching of double-encrypted IDs. Now, both publisher and advertiser would hold a dataset with only matched customers and their accompanying attributes but both parties would not know the identities of the common customers as the IDs are doubly encrypted.



id	[conv_ts]	[val]
*	[9]	[0]
g	[9]	[4]
c	[8]	[7]
e	[4]	[5]

id	[impr_ts]
*	[5]
g	[3]
c	[5]
e	[3]

Matched identifiers with accompanying attributes after PSI

Private Set Intersection

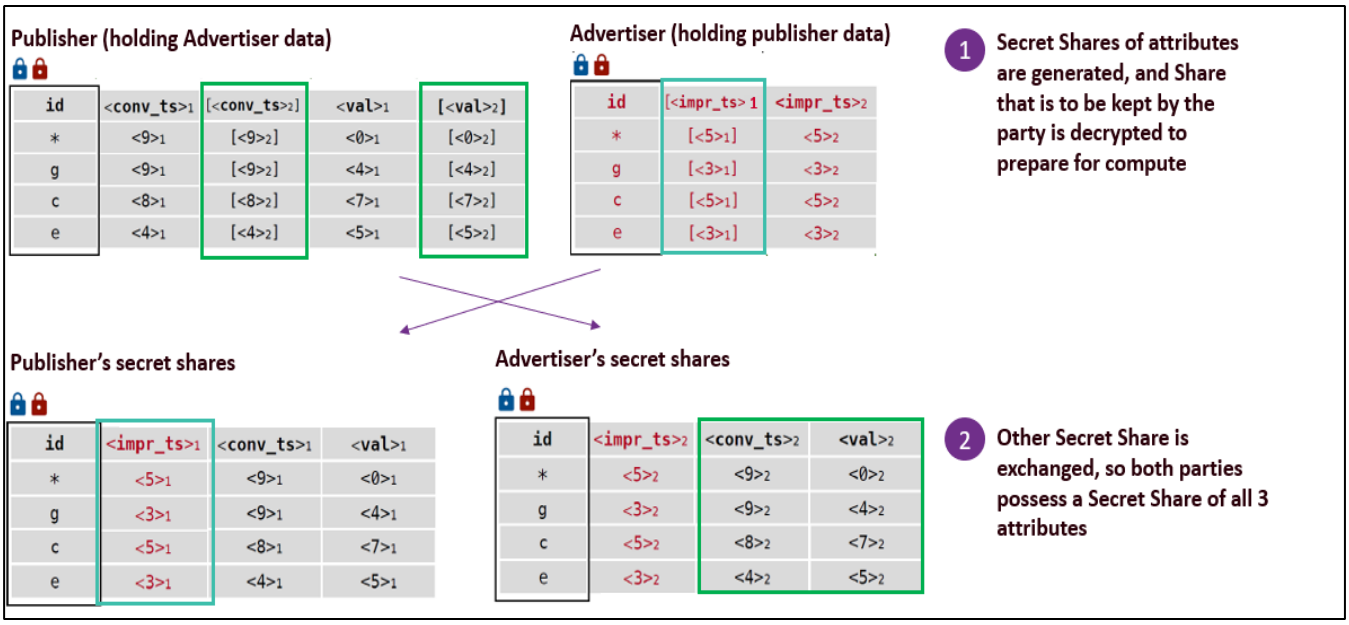
a. The Advertiser and TikTok use PSI to identify the list of common customers in both their datasets without revealing their data or exposing information on whether any of their customers are common customers or not.

b. The PSI implementation uses elliptic curve cryptography (“ECC”)[1] to encrypt their customer IDs with both parties’ private key (double encryption) and identifying the matched records, i.e records with same doubly encrypted customer ID values.

ECC is the main enabler of PSI through the hiding of plaintext customer ID from being revealed. The ECC implementation uses the open-source OpenSSL[2] libraries that are commonly used by many programs (e.g. Google Chrome) when implementing cryptographic operations. The ECC curve chosen in this implementation is P-256 and defined in NIST standards SP 800-186.

Phase 2 – Computation of conversion values

e. Step 4 – The encrypted accompanying attributes (i.e. conversion timestamp, conversion value and impression timestamp) are “shredded” into 2 secret shares and publisher and advertisers will each hold one secret share of all 3 attributes. *Please see diagram below*



Publisher (holding Advertiser data)

id	<conv_ts>$_1$	[<conv_ts>$_2$]	<val>$_1$	[<val>$_2$]
*	<9>$_1$	[<9>$_2$]	<0>$_1$	[<0>$_2$]
g	<9>$_1$	[<9>$_2$]	<4>$_1$	[<4>$_2$]
c	<8>$_1$	[<8>$_2$]	<7>$_1$	[<7>$_2$]
e	<4>$_1$	[<4>$_2$]	<5>$_1$	[<5>$_2$]

Advertiser (holding publisher data)

id	[<impr_ts> 1	<impr_ts>$_2$
*	[<5>$_1$]	<5>$_2$
g	[<3>$_1$]	<3>$_2$
c	[<5>$_1$]	<5>$_2$
e	[<3>$_1$]	<3>$_2$

1 Secret Shares of attributes are generated, and Share that is to be kept by the party is decrypted to prepare for compute

Publisher’s secret shares

id	<impr_ts>$_1$	<conv_ts>$_1$	<val>$_1$
*	<5>$_1$	<9>$_1$	<0>$_1$
g	<3>$_1$	<9>$_1$	<4>$_1$
c	<5>$_1$	<8>$_1$	<7>$_1$
e	<3>$_1$	<4>$_1$	<5>$_1$

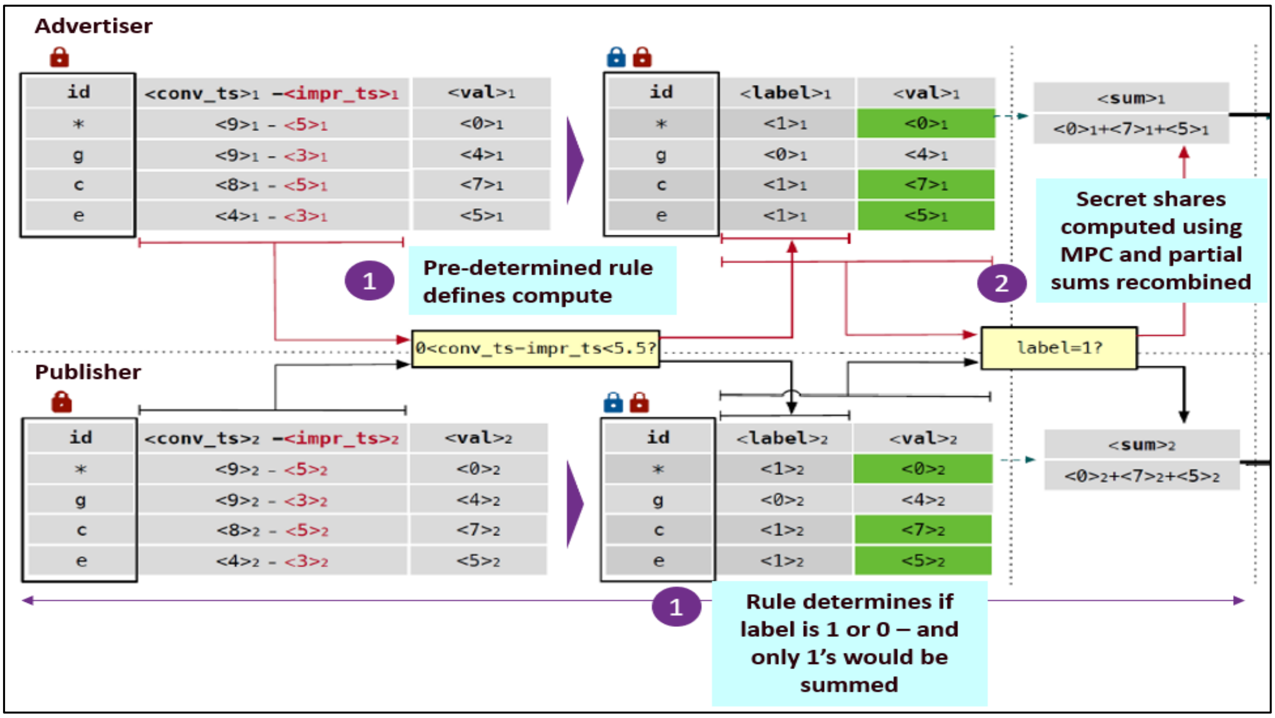
Advertiser’s secret shares

id	<impr_ts>$_2$	<conv_ts>$_2$	<val>$_2$
*	<5>$_2$	<9>$_2$	<0>$_2$
g	<3>$_2$	<9>$_2$	<4>$_2$
c	<5>$_2$	<8>$_2$	<7>$_2$
e	<3>$_2$	<4>$_2$	<5>$_2$

2 Other Secret Share is exchanged, so both parties possess a Secret Share of all 3 attributes

Publisher and Advertiser holding encrypted secret shares

f. **Step 5** – Based on pre-determined criteria between publisher and advertiser, (i.e. time between impression and conversion is < 5 seconds), the ad conversion computation then takes place using Secure Multi-Party Computing (SMPC) additive secret sharing technique to hide the computation from both parties. The two main steps of the computation involve 1) selecting records by determining if the difference of conversion time against the impression time is <5s, and 2) sum all selected records' conversion values to determine the total conversion value for the ad.



Computation of ESS based on pre-determined rules

<u>Secure Multi-Party Computation</u>

a. The matched customers' conversion values are in secret shares (SMPC) from the previous HE operation. Secure Multi-Party Computation protocol ("SMPC")[5] is used to protect the ad conversion values of each customer during sharing and computation thereafter.

b. The ad conversion computation then takes place using SMPC additive secret sharing technique through the ABY protocol[2] and CrypTFlow2[3] protocol to hide the computation from both parties.

<u>Differential Privacy</u>

a. During the computation of the aggregated conversion value through SMPC, Differential Privacy (DP), i.e. noise, is added through the MPC-Dual DP protocol. This is accomplished through an Oblivious Transfer ("OT") protocol[9] to mathematically distribute noise between the parties for addition to their respective secret share values.

[2] ABY Protocol – A Framework for Efficient Mixed-Protocol Secure Two-Party Computation

[3] CrypTFlow2: Practical 2-Party Secure Inference

The use of DP further protects the final aggregated conversion value and mathematically guarantees to reduce the risk of re-identification and inference of the conversion values to any individual.

Regulatory Learnings

9. TikTok sought Practical Guidance (“PG”) from the Personal Data Protection Commission (PDPC) on whether the data shared between the Advertiser and Publisher in the POC constitutes personal data such that the data protection obligations under the Personal Data Protection Act 2012 (“PDPA”) would apply.

10. Personal data is defined in Section 2 of the PDPA to refer to data, whether true or not, about an individual who can be identified (a) from that data; or (b) from that data and other information to which the organization has or is likely to have access.

11. PDPC is of the view that in this POC, there is **no serious possibility that each party will be able to identify individuals from the data that it receives from the other party, or that a third party who obtains the shared data will be able to identify individuals from it.** Hence, the data sharing between the parties does not constitute a disclosure or collection of personal data under the PDPA for the disclosing party and the receiving party respectively. In coming to this view, the PDPC considered the following:

    a. The adding of “noise” by adding dummy records and shuffling records;
    b. The use of effective encryption techniques to prevent recipients from gaining access to information or insights relating to the common customers;
    c. That each party keeps its encryption keys and secret share “shred” confidential from the other party and any other persons; and
    d. That the final output is an aggregated conversion value that is obtained through Secure Multi-Party Computing (SMPC) where neither party would be able to attribute or link to any identifiable individual.

12. TikTok and the Advertiser should ensure that the measures and security arrangements implemented to prevent the risk of re-identification and protect against data protection threats remain effective and up-to-date. This includes keeping the PETs used in the POC updated with prevailing industry-recognised processes and standards and ensuring “cryptographic agility” by replacing cryptography algorithms that are found to be vulnerable.

Results and Next Steps

13. **The overall performance and time taken for the computation of PrivacyGo** were **deemed acceptable for ad measurement** for the largest datasets (50M) tested, end-to-end computation took about 10 hours to complete. The timeframe was acceptable as ad measurement was not time-sensitive. Significantly longer processing times—for example, days or even weeks—would render the solution unfeasible. The delays would be disruptive and impact the business's ability to act on insights in a timely manner.

14. The PoC results demonstrated that the precision of computations from the **PPAM protocol matched the accuracy of non-encrypted computations up to two decimal places.** (see Annex 1) This ability to produce accurate measurements, even with the addition of a PET, aligned well with advertisers' needs to balance privacy vs. utility.

15. PrivacyGo demonstrates an effective and innovative fusion of multiple PETs, showing that with focused engineering effort, it is possible to build protocols that are practical for real-world use while meeting key privacy requirements.

16. Nevertheless, whilst **performance was adequate for practical deployment**, there remains room for further improvement. The current MPC implementation of PrivacyGo is single-threaded[4]; to improve performance, multi-threading could be explored to process and compute data in parallel, possibly reducing overall execution time

[4] "Single-threading" is the processing of one instruction at a time, in a sequential manner as opposed to multi-threading which has multiple operations happening concurrently.

Annex 1 – End to end testing results

	TikTok's data size	Advertiser's data size	Intersection size			Total conversion value		Communication(MB)	Time (sec)
			PII 1	PII 2	Total	Non-encrypted computation	AWS E2E Testing Results	AWS E2E Testing Results	AWS E2E Testing Results
Testing Data sets	10M	100k	8000	1589	9589	358601.83	358601.8295	6385.51	7302.85
	20M	200k	16000	3203	19203	729127.81	729127.8106	12755	14548.3
	30M	300k	24000	4808	28808	1089432.56	1089432.5602	19124.5	21920.7
	40M	400k	32000	6333	38333	1454868.7	1454868.6997	25493.8	29167.2
	50M	500k	40000	7999	47999	1811168.38	1811168.3809	31879.4	36525.8