Guessing Isn't Good Enough: Privacy-Preserving Metadata Collection in Mobile Messaging

Alexandre Pauwels
University of Cambridge
ap2453@cam.ac.uk

Alastair R. Beresford
University of Cambridge
arb33@cam.ac.uk

***Abstract*—Metadata collection on secure mobile instant messaging platforms remains controversial. Academics, governments, law enforcement, and health departments have long argued it could be used to aid in such commendable endeavours as preventing the spread of misinformation, catching criminals, and reducing abuse. Privacy advocates, on the other hand, warn that disingenuous analysis of the same data by bad actors could be used to unfairly target individuals or justify harmful decisions like the banning of end-to-end encryption entirely. Thus far, the most secure mobile messaging platforms, like Signal, have self-imposed a strict zero-tolerance policy on metadata collection. Recent work, however, has shown that understanding mobile messaging behaviour is a requirement for relatively uncontroversial technical decision-making like selection of an optimally secure post-quantum communications protocol. We propose a framework using local differential privacy that allows mobile messaging platforms to begin making such decisions without revealing information about any individual user. Our results show that population-level queries can be usefully answered with a user-level lifetime privacy budget of $\varepsilon = 1$ using as few as 1.8 million participants, which is small relative to platforms the size of WhatsApp, Signal, and iMessage.**

***Index Terms*—local differential privacy, mobile, messaging, signal, whatsapp, imessage**

I. Introduction

The metadata generated by daily usage of mobile messaging platforms represents a double-edged sword of utility and danger. On one hand, metadata analysis can help curb the spread of health misinformation [1], identify terrorists [2], and prevent suicide [3]. On the other, it can unfairly accuse users of possessing child sexual-abuse material (CSAM) [4] and help proponents of mass surveillance justify banning access to strong encryption [5]. To prevent any potential for abuse, mobile messengers like Signal minimize metadata collection by building communications protocols that hide metadata when possible and avoid logging it otherwise [6].

Not all queries that could be answered by metadata analysis have the potential for evil, however. Recent work in post-quantum secure messaging shows that selection of an optimally secure protocol depends on the density and frequency of message exchange within a conversation [7]. Recent papers by Google [8] and Oratomic [9] have compressed the timeline for migration to post-quantum secure cryptography [10], creating an urgent need for validation of protocol security on real conversations. Similar work could be performed on CoverDrop, a metadata-private secure messaging system for whistleblowers built into The Guardian's mobile news app [11]. CoverDrop employs traffic shaping strategies which hide whether the feature is used at all from a traffic analysing adversary. Since usage of the feature is itself sensitive metadata, CoverDrop operators are unable to measure telemetry indicating the messaging latency experienced by users, and which parts of the systems could be improved to safely reduce that latency. These two examples show that there exists a category of queries that are both non-evil and necessary but cannot currently be answered.

We propose using local differential privacy (LDP) to allow secure mobile messaging platforms to answer these queries while respecting user privacy. Rather than submitting raw metadata, users first independently noise their submission such that it no longer directly reflects their state. Noise is carefully calibrated such that the server can combine the noised data and, given sufficient submissions, extract accurate population-level statistics. Thanks to the size of popular messaging platforms, which range from tens of millions to billions of users, the magnitude of the noise can be such that the data submitted from individual users is essentially nonsense while still being useful in aggregate. In contrast to other popular LDP solutions deployed at scale by Google [12] and Apple [13], our framework has users assign themselves a privacy budget which gets subtracted from on each submission. Once that budget reaches zero, the user no longer contributes any noised metadata. This provides an incentive for platforms to carefully select queries that are both worthy of and minimise their population-level privacy spend.

We evaluate our framework by estimating the number of users and privacy-spend per user required to select an optimal post-quantum secure messaging protocol. Users simulate each protocol being compared on their conversations, and for each produce a set of histograms which reflect the security of that protocol. Independent Laplace noise is added to each bin, and bins are averaged at the server to recover population-level statistics. We show that, with 1.8 million participating users, our methodology can identify which protocol leads to the greatest population-level security while consuming a privacy budget of $\varepsilon = 1$ per user. Increasing the number of participating users can reduce the spend or increase accuracy.

REFERENCES

[1] V. Papanikou, P. Papadakos, T. Karamanidou, T. G. Stavropoulos, E. Pitoura, and P. Tsaparas, "Health Misinformation in Social Networks: A Survey of Information Technology Approaches," *Future Internet*, vol. 17, p. 129, Mar. 2025.

[2] "Report: Terrorist Use of End-to-End Encryption - Insights from a Year of Multi-Stakeholder Discussion." https://techagainstterrorism.org/news/2023/01/11/terrorist-use-of-end-to-end-encryption-insights-from-a-year-of-multi-stakeholder-discussion.

[3] J. J. Glenn, A. L. Nobles, L. E. Barnes, and B. A. Teachman, "Can Text Messages Identify Suicide Risk in Real Time? A Within-Subjects Pilot Examination of Temporally Sensitive Markers of Suicide Risk," *Clinical Psychological Science*, vol. 8, pp. 704–722, July 2020.

[4] S. Klier and H. Baier, "To Possess or Not to Possess - WhatsApp for Android Revisited with a Focus on Stickers," in *Secure IT Systems: 28th Nordic Conference, NordSec 2023, Oslo, Norway, November 16–17, 2023, Proceedings*, (Berlin, Heidelberg), pp. 281–303, Springer-Verlag, Nov. 2023.

[5] R. Anderson, "Chat Control or Child Protection?." https://arxiv.org/abs/2210.08958v1, Oct. 2022.

[6] "Signal is for everyone, and everyone is different." https://signal.org/blog/signal-is-for-everyone/.

[7] B. Auerbach, Y. Dodis, D. Jost, S. Katsumata, and R. Schmidt, "How to Compare Bandwidth Constrained Two-Party Secure Messaging Protocols: A Quest for A More Efficient and Secure Post-Quantum Protocol," in *34th USENIX Security Symposium (USENIX Security 25)*, pp. 6717–6736, 2025.

[8] R. Babbush, A. Zalcman, C. Gidney, M. Broughton, T. Khattar, H. Neven, T. Bergamaschi, J. Drake, and D. Boneh, "Securing Elliptic Curve Cryptocurrencies against Quantum Vulnerabilities: Resource Estimates and Mitigations," Apr. 2026.

[9] M. Cain, Q. Xu, R. King, L. R. B. Picard, H. Levine, M. Endres, J. Preskill, H.-Y. Huang, and D. Bluvstein, "Shor's algorithm is possible with as few as 10,000 reconfigurable atomic qubits," Mar. 2026.

[10] "Quantum frontiers may be closer than they appear." https://blog.google/innovation-and-ai/technology/safety-security/cryptography-migration-timeline/, Mar. 2026.

[11] D. Hugenroth, S. Cutler, D. Kendrick, M. Savarese, Z. Hunter-Green, P. McMahon, M. Kalanaki, D. A. Vasile, S. Bejasa-Dimmock, L. Hoyland, and A. R. Beresford, "CoverDrop White Paper," tech. rep., Computer Laboratory, University of Cambridge.

[12] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, (Scottsdale Arizona USA), pp. 1054–1067, ACM, Nov. 2014.

[13] "Learning with Privacy at Scale." https://machinelearning.apple.com/research/learning-with-privacy-at-scale.