Infer User Preferences from Aggregate Measurements: A Novel Message Passing Algorithm for Privacy Attack

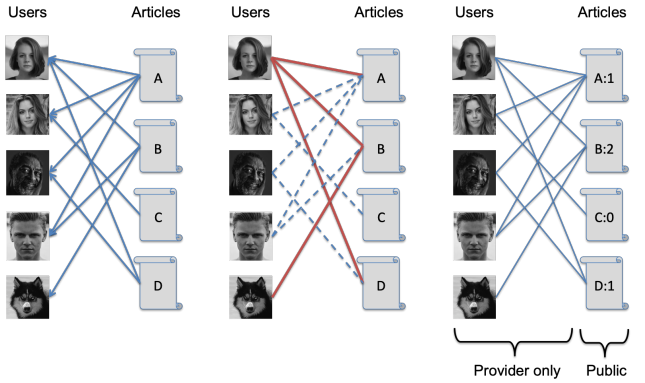
[extended abstract]

Du Su , Yi Lu
University of Illinois at Urbana-Champaign
{dusu3, yilu4}@illinois.edu

Social media platforms, such as Facebook and TikTok, have triggered debates on privacy. The recent transformation of social media into an increasingly centralized service, exemplified by TikTok, only exacerbates the matter. While aggregation has been deemed an effective way to combat privacy infringement, a high degree of centralization can make aggregation ineffective.

We present a randomized push algorithm, with which a service provider can infer individual users' preferences from publicly available aggregate data. Hence, even social media platforms comply with strict privacy regulations regarding individual user's action, a tremendous amount of information can still be inferred from aggregate data due to the centralized control.

In order to infer users' individual preferences, the social-platform-service provider chooses a set of contents for each topic of interest. The topic of a piece of content is known to the service provider. Figure 1 illustrates the push algorithm for one topic. In Fig. 1(a), each content is pushed to a random subset of users. In Fig. 1(b), if a user is interested in the topic, he views the pushed contents; otherwise, he does not view them. However, the individual action of a user viewing a particular piece of content is not recorded in the system. Instead, only the *aggregate* number of views of each content, as in Fig. 1(c), is recorded.



(a) Each content is pushed to a random subset of users.

(b) Interested users view the pushed contents. This individual action is not recorded in the system.

(c) Aggregated views for each article is recorded and is publicly available, but the push-algorithm is only known to the service provider.

Figure 1: Overview of the push-aggregation process.

1. PROBLEM FORMULATION

Given a set of m articles of a given topic and a set of n users whose preferences of the topic are denoted by $p_i \in \{0,1\}$, $i = 1, \cdots, n$, each user is pushed a subset of the articles. The aggregate views of each article, $r_a = \sum_{i \in I(a)} p_i$, $a = 1, \cdots, m$, are collected, where $I(a)$ is the set of users whom article a is pushed to. The inference problem is to find the values of p_i, for all $i = 1, \cdots, n$.

We are interested in designing a push algorithm and an inference algorithm that achieve the following: 1) The push algorithm only sends a small number of articles to each user. 2) The inference algorithm is of low complexity. 3) Together the algorithms find the user preferences with as *few* articles as possible.

1.1 Formulation as a Compressed Sensing Problem

The above problem can also be formulated as a class of compressed sensing problem [Keiper et al. 2017], where a signal $x \in \{0,1\}^n$ is recovered from a underdetermined system of linear equations

$$Ax = b$$

. Our problem has the following additional constraints motivated by the application: 1) Matrix A in a compressed sensing problem has real-number entries. However, for our problem the entries of A have to be in $\{0,1\}$, denoting whether an article is pushed to a user. 2) Matrix A in a compressed sensing problem does not have sparsity requirement. However each user is only able to read a small number of articles, hence A has to be *very sparse*. 3) The reconstruction algorithms for compressed sensing has a worst-case $O(n^3)$ complexity. With n being large in our application, we are interested in designing a *linear-complexity* inference algorithm.

2. CONTRIBUTIONS

We propose the following solution that consists of two parts:

1. A randomized push algorithm that sends only $O(1)$

articles to each user.

2. A message-passing reconstruction algorithm that infers p_i from the aggregate views r_a. The complexity of the message passing algorithm is $O(n)$, i.e., linear in the number of users.

The push algorithm can be formulated as a bipartite graph between users and articles, where the article node a is connected to $I(a)$. We describe the bipartite graph by its degree distribution:

DEFINITION 1. **Degree distribution.** *Given a bipartite graph with n user nodes and m article nodes, let λ_k be the fraction of edges that connect a user node with degree k, and ρ_k be the fraction of edges that connect an article nodes with degree k, then the edge-perspective degree distribution pair is*

$$\lambda(x) = \sum_{k=1}^{l_{max}} \lambda_k x^{k-1}, \qquad \rho(x) = \sum_{k=1}^{r_{max}} \rho_k x^{k-1}$$

The push algorithm can be designed to minimize the number of articles needed for accurate reconstruction with probability 1. Let the proportion of users who prefer the given topic be $\epsilon \in [0,1]$. Let the average number of articles per user, $\beta = \frac{m}{n}$. We study the threshold β^* such that all for all $\beta > \beta^*$, the preferences p_i can be inferred correctly with probability 1.

We obtain the following results:

THEOREM 1. **Optimal Ratio.** *The optimal article-to-user ratio β^* satisfies the following bounds:*

$$\beta^* \leq \sqrt{\epsilon(1-\epsilon)}.$$

THEOREM 2. **Achievability.** *We can construct a sequence of degree distribution pairs, $(\lambda_\alpha^{(N)}, \rho_\alpha)$ with*

$$\lim_{N\to\infty} \beta^{(N)} \triangleq \lim_{N\to\infty} \frac{\int_0^1 \rho_\alpha(x)dx}{\int_0^1 \lambda_\alpha^{(N)}(x)dx} = \sqrt{\epsilon(1-\epsilon)}$$

such that in the large n limit, all user preferences can be correctly reconstructed with probability 1.

THEOREM 3. **Phase Transition.** *In the large n limit, there exists a threshold of the proportion of interested users ϵ^*, such that:*
(i) If $\epsilon \leq \epsilon^*$, $\hat{p}_i(2t) \uparrow p_i$ and $\hat{p}_i(2t+1) \downarrow p_i$ for all i.
(ii) If $\epsilon > \epsilon^*$, there exists a positive proportion of users such that $\hat{p}_i(2t) < \hat{p}_i(2t+1)$ for all t. Thus, some users' preferences are not correctly inferred.

Note that the threshold β^* is the exact counterpart of the phase-transition threshold analyzed in [Keiper et al. 2017] where the compressed sensing problem with the signal $x \in \{0,1\}^n$ is solved with the basis pursuit algorithm. The result in [Keiper et al. 2017] is on random Gaussian matrices and the threshold can be numerically obtained. We plot the compressed sensing threshold, denoted by CS, against our threshold in Figure 2.

Comparison to Counter Braids [Lu et al. 2008]

We also compare our algorithm against that of one-layer Counter Braids[Lu et al. 2008] (CB) in Figure 2. The CB algorithm is designed for an infinite alphabet of $x \in \mathbb{N}$ whereas our algorithm is designed for $x \in \{0,1\}$. We observe that

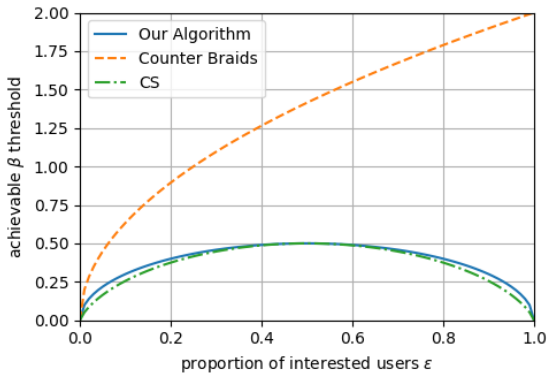


Figure 2: The proposed algorithm achieves a phase-transition threshold that is extremely close to that of compressed sensing (CS) and significantly lower than one-layer Counter Braids.

phase transition occurs much earlier with our algorithm than with the CB algorithm.

The analysis of our reconstruction algorithm deals with the *lack of monotonicity* in ϵ and can also be of interest to general applications involving compressed measurement. Note that the lack of monotonicity is *not* due to the symmetry with respect to $\epsilon = 0.5$, but is inherent in the message passing algorithm even if we restrict to $\epsilon < 0.5$.

In particular, given a fixed pair of push and reconstruction algorithms, the density evolution equation of both the erasure-decoding algorithm [Richardson and Urbanke 2008] and Counter Braids [Lu et al. 2008], has the following property: for any $\epsilon_1 < \epsilon_2$, the corresponding error probabilities at the t-th iteration always satisfy $P_1^t < P_2^t$, i.e., the monotonicity of error probability is kept at each iteration.

With our algorithm, however, given $\epsilon_1 < \epsilon_2 < 0.5$, the monotonicity of $P_1^t < P_2^t$ does *not* hold due to the asymmetry inherent in the message passing algorithm. It is hence surprising that phase transition occurs with the degree distribution achieving $\beta = \sqrt{\epsilon(1-\epsilon)}$, despite the lack of monotonicity.

3. REFERENCES

S. Keiper, G. Kutyniok, D. G. Lee, and G. E. Pfander. 2017. Compressed Sensing for Finite-Valued Signals. *Linear Algebra Appl.* 532 (2017), 570–613.

Yi Lu, Andrea Montanari, Balaji Prabhakar, Sarang Dharmapurikar, and Abdul Kabbani. 2008. Counter braids: a novel counter architecture for per-flow measurement. *ACM SIGMETRICS Performance Evaluation Review* 36, 1 (2008), 121–132.

Tom Richardson and Ruediger Urbanke. 2008. *Modern Coding Theory*. Cambridge University Press, USA.