Measures of cross-site re-identification risk: An analysis of the Topics API Proposal

Alessandro Epasto
Google Research

Andres Muñoz Medina
Google Research

Christina Ilvento
Chrome

Josh Karlin
Chrome

Abstract

We present an analysis of the ability of a third party to use the Topics API to re-identify a user across two different sites. Our analysis explores how some of the design choices affect this capability. We present both worst-case and optimistic scenarios for user re-identification, and provide empirical evidence that the topics revealed are closer to the optimistic scenario than the worst-case one.

1 Introduction

Several API proposals have been put forward as part of the Privacy Sandbox effort; these proposals are intended to significantly reduce cross site tracking while preserving relevant advertising. In comparison to third party cookies, these APIs represent a step forward towards improving user privacy by supporting several key ads use cases without using persistent cross-site identifiers. However, any API that provides information about a user's interests based on past browsing behavior does reveal some cross-site information. In this note, we seek to analyze and quantify the information leaked by the proposed Topics API and to show how different design choices impact its privacy properties.

We present a formal model to analyze privacy leakage in the Topics API. The approach allows us to precisely state the worst case risk of re-identification and to identify the key properties of the real world data distribution that ameliorate some of the worst case bounds. We evaluate the risks both analytically and using real world empirical data. We make several assumptions on the threat model and we discuss limitations of those assumptions in Section 7.

2 Setup

As previous analyses have noted [3], establishing a threat model for the Topics API requires care, since the API is designed to provide some cross-site interest information about user behavior to support advertising. Our goal is to consider re-identifiability: roughly speaking, the ability of an attacker using the Topics API to link the identity of a user across two different domains.

We now define this formally. Consider a universe $\mathcal{U}$ of n users that visit websites. We denote a user by u, and websites by $w \in \mathcal{W}$. The space $\mathcal{W}$ consists of the collection of all domains in the internet. Abstractly, the Topics API can be modeled as a randomized function $\mathsf{T}\colon \mathcal{W} \times \mathcal{U} \to \mathcal{C}$ where $\mathcal{C}$ is the set of possible categories that can be returned by the API.

An attacker is able to observe the topics returned by the API on two websites[1], w_1 and w_2. For a user u, let $\mathsf{i}_1(u)$ and $\mathsf{i}_2(u)$ represent the identity of user u on w_1 and w_2 respectively. We assume $\mathsf{i}_1(u)$, $\mathsf{i}_2(u)$ are the ids assigned to the user u by the two sites, and these ids are uncorrelated with each other. We assume the ids of the users of a site are unique. The attacker sees all n users $u \in \mathcal{U}$ visit both sites, generating sets of

[1]This stylized model makes several simplifying assumptions to provide a more focused threat model and analysis. In practice, the attacker will likely be able to observe topics on multiple websites, and the output of the topics API may be derived from more than one site visit.

tuples $\mathfrak{T}_1 = \{(\mathsf{i}_1(u), \mathsf{T}(w_1, u))\}_{u\in\mathcal{U}}$ and $\mathfrak{T}_2 = \{(\mathsf{i}_2(u), \mathsf{T}(w_2, u))\}_{u\in\mathcal{U}}$. Nature picks a user $U \in \mathcal{U}$ uniformly at random, and reveals to the attacker the corresponding pair $(\mathsf{i}_2(U), \mathsf{T}(w_2, U))$ on website w_2. The attacker's goal is to select the identity of the random user on website w_1, i.e. $\mathsf{i}_1(U)$.

There are two threat models that we consider:

- Per-Instant: The attacker sees a single output from the topics API for each user.
- Longitudinal: The attacker has access to consistent first party identity, and thus sees a sequence of topics corresponding to each user.

2.1 Examples

Third party cookies We can model third party cookies as $\mathsf{T}(w, u) = h(u)$ where h is a random hash function with a large range. Observe that since h does not depend on w, the attacker's job is trivial in both the per-instant and longitudinal threat models.[2]

Topics API We model the proposed Topics API by parameterizing it in three ways: the total number of topics, $|\mathcal{C}| = N$, the number of topics being randomized amongst, k, and the probability of selecting a random topic p.

For $k > 0$ and user $u \in \mathcal{U}$ let $S := S(u, k) \subset \mathcal{C}$ denote the top-k categories associated with a user's browsing history in a given epoch. In this document we make no assumption on the set $S(u, k)$ other than $|S(u, k)| = k$. The latter can always be ensured by padding the top-k categories with random topics whenever the user has fewer than k categories. The Topics API tracks $S(u, k)$ for each user u, and the function $\mathsf{T}(w_1, u)$ returns one of the elements of $S(u, k)$ uniformly at random with probability $(1 - p)$ and a uniformly at random topic from $\mathcal{C}$ with probability p. Notice that the topic $T(w_1, u)$ is fixed for every visit of u to site w_1 for a given epoch of the API.

Throughout the paper we assume that the sets $S(u, k)$ are fixed yet unknown to the attacker. The only access to these sets is through the Topics API.

3 Re-identification Metrics

An immediate measure of the success of the attacker is the probability (over the random choice of user) that the attack is successful.

Formally, let U denote a uniform random variable over the set of users $\mathcal{U}$ and let $T_2 = \mathsf{T}(w_2, U)$ be the random variable representing the output of the API on w_2. Without any additional information, an attacker can only guess the identity of U in w_1 uniformly at random. That is, the prior of the user's identity in w_1 is given by:

$$P(\mathsf{i}_1(U) = i) = \frac{1}{n}.$$

As described before, the attacker also has access to the set $\mathfrak{T}_1 = \{(\mathsf{i}_1(u), \mathsf{T}(w_1, u)) \colon u \in U\}$, i.e., all topics observed on site w_1, and the random variable T_2. Notice that $\mathfrak{T}_1$ is a *random* set as the Topics API samples an element from the top set of a user's topics. We will denote realization of this random set by $\mathfrak{t}_1$.

Remark 1. *Throughout the document the probability measure we are considering is induced only by the random selection of a user U and the randomness of the function T itself. The actual set of topics for each user, $S(u, k)$ is fixed, yet unknown to the attacker.*

[2]Recall that the goal of the Topics proposal is to provide relevant advertising while significantly reducing cross-site tracking compared with third-party cookies. This formulation demonstrates that third-party cookies have arbitrary leakage of cross-site information, and that by comparison, the Topics API has a bounded rate of leakage. The aim of our analysis is to understand how meaningful this bound is in practice.

Given a realization of T_2 and $\mathfrak{T}_1$, the attacker can obtain a better posterior on the identity of U. In particular, they may adjust their belief as

$$P(\mathfrak{i}_1(U) = i | \mathfrak{T}_1 = \mathfrak{t}_1, T_2 = t_2) \tag{1}$$

Observe that if this distribution is highly concentrated on a single identity value, then the re-identification attack can be successful.

To precisely measure the information learned about the user by revealing a topic on w_2 we look at the KL-divergence between this distribution and the uniform prior:

$$\mathrm{KL}\Big(P(\mathfrak{i}_1(U) = \cdot | \mathfrak{T}_1 = \mathfrak{t}_1, T_2 = t_2) \;||\; P(\mathfrak{i}(U) = \cdot)\Big).$$

This is precisely the information gain on $\mathfrak{i}_1(U)$ due to learning that $T_2 = t_2$ and $\mathfrak{T}_1 = \mathfrak{t}_1$.

When taking the expectation of the above quantity with respect to both the choice of t_2 and $\mathfrak{t}_1$ we recover a well studied concept in information theory: the mutual information [1] between a user's identity $\mathfrak{i}_1(U)$ and the topics revealed by the Topics API:

$$I(\mathfrak{i}_1(U); \mathfrak{T}_1, T_2) = \mathbb{E}_{t_2, \mathfrak{t}_1} \mathrm{KL}\Big(P(\mathfrak{i}_1(U) = \cdot | \mathfrak{T}_1 = \mathfrak{t}_1, T_2 = t_2) \;||\; P(\mathfrak{i}(U) = \cdot)\Big).$$

The above term represents the average number of bits of information that T_2 and $\mathfrak{T}_1$ provide about user's identity on w_1.

Before analyzing this quantity for the Topics API, we first simplify this expression.

Theorem 1. *Let $T_1 = \mathsf{T}(w_1, U)$. The above mutual information can be simplified as:*

$$I(\mathfrak{i}_1(U); \mathfrak{T}_1, T_2) = I(T_1; T_2),$$

where $I(T_1; T_2)$ denotes the mutual information between the random variables T_1 and T_2.

Intuitively, if observing T_2 allows an attacker to predict the topic of the user on w_1 then an attacker can reduce their search on w_1 to those users whose topic matches the prediction.

4 Per Instance Analysis

We begin with a worst case analysis for the number of bits leaked by the Topics API in a single time step. It is not hard to see that the mutual information depends on the collection of top topics $\mathcal{S} = \{S(u_1, k), \ldots, S(u_n, k)\}$. We call a collection of top k sets an assignment of topics to users. Thus we want to understand how much leakage can happen across all possible assignments. Each assignment $\mathcal{S}$ induces a probability measure P through the random variable $S(U, k)$. We denote by $\mathcal{P}$ the finite set of all such probability measures.

Theorem 2. *Let $N = |\mathcal{C}|$ be the number of topics in a taxonomy, and let $\mathcal{P}$ be defined as above, then*

$$\max_{P \in \mathcal{P}} I(T_1; T_2) \leq \log \frac{N}{k}.$$

Proof. Using the definition of mutual information we have:

$$I(T_1; T_2) = H(T_2) - H(T_2 | T_1),$$

where $H(\cdot)$ and $H(\cdot|\cdot)$ denote the entropy and conditional entropy operators. The first term is clearly upper bounded by $\log N$. We now lower bound the second term. Let $S = S(U, k)$, using the fact that conditioning always decreases entropy we have:

$$H(T_2 | T_1) \geq H(T_2 | T_1, S) = H(T_2 | S),$$

where the last equality holds since the distribution of T_2 does not depend on the first topic once S is known. Finally since T_2 is chosen uniformly from S and $|S| = k$ we must have $H(T_2|S) = \log k$. □

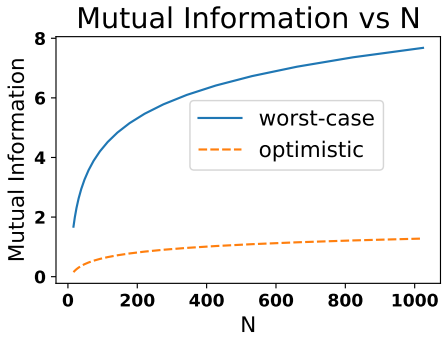

Figure 1: Mutual information in the optimistic and worst case scenario.

In other words, the Topics API leaks at most $\log {}^{N}/_{k}$ bits with every topic release. For $N = 349$ – the size of the published taxonomy [2] – and $k = 5$, this is 6.12 bits. It is worth noticing that even in the worst case scenario this is still less cross-site information leakage than expected. Indeed, at first glance, the Topics API seems to provide $\log N \sim 8.44$ bits of cross-site tracking information. The above theorem shows that we can reduce this information by 26% even in the worst case scenario.

The following example shows that the above bound can be achieved.

Example 1 (Worst case scenario)**.** *Let $N = |\mathcal{C}|$ denote the number of topics in $\mathcal{C}$ and assume that N is divisible by k. Let $\{P_1, \ldots, P_{N/k}\}$ denote a partition of $\mathcal{C}$. We refer to each element in the partition as a topic profile. Let the top set $S = S(U, k)$ follow a uniform distribution over the profile partition. We proceed to calculate the mutual information between T_1 and T_2.*

$$I(T_1; T_2) = H(T_1) - H(T_1|T_2)$$

It is clear that T_1 is distributed uniformly across all possible topics. We thus have $H(T_1) = \log N$. On the other hand, because S is a topic profile, observing T_2 fully determines the value of S. Therefore, conditioned on T_2, T_1 can be only one of k possible values. It follows that $H(T_1|T_2) = \log k$ and

$$I(T_1; T_2) = \log \frac{N}{k}$$

Observe that this example requires a very precise and clustered distribution of top-k topics. On the other hand, if all sets of top-k are equally likely, then the number of bits leaked is significantly smaller.

Example 2 (Optimistic scenario)**.** *Consider the Topics API with $p = 0$. Assume that the random variable $S = S(U, k)$ is distributed uniformly across all possible length k sequences of topics. We know that*

$$P(T_1 = t_1, T_2 = t_2) = P(T_1 = t_1, T_2 = t_2 | (t_1, t_2) \in S) P((t_1, t_2) \in S) \tag{2}$$

The first term is always equal to $\frac{1}{k^2}$ whereas, due the uniformity of S, the second is given by

$$\begin{cases} \frac{k}{N} & t_1 = t_2 \\ \frac{k(k-1)}{N(N-1)} & t_1 \neq t_2 \end{cases}$$

Therefore

$$P(T_1 = t_1, T_2 = t_2) = \begin{cases} \frac{1}{kN} & t_1 = t_2 \\ \frac{(k-1)}{kN(N-1)} & t_1 \neq t_2 \end{cases}$$

$$
\begin{aligned}
I(T_1;T_2) &= \sum_{t\in\mathcal{C}} P(T_1=T_2=t)\log\frac{P(T_1=T_2=t)}{P(T_1=t)P(T_2=t)} + \sum_{t_1\neq t_2} P(T_1=t_1,T_2=t_2)\log\frac{P(T_1=t_1,T_2=t_2)}{P(T_1=t_1)P(T_2=t_2)} \\
&= \sum_{t\in\mathcal{C}} \frac{1}{kN}\log\frac{\frac{1}{kN}}{\frac{1}{N^2}} + \sum_{t_1\neq t_2}\frac{k-1}{kN(N-1)}\log\frac{(k-1)N}{(N-1)k} \\
&= \frac{1}{k}\log\frac{N}{k} + \frac{k-1}{k}\log\frac{N(k-1)}{k(N-1)}
\end{aligned}
$$

For $k=5$ and $N=349$ the above expression is approximately 0.97 bits which is considerably less than the 6.12 bits in the worst case. We report this bound and the worst-case bound for $k=5$ and different values of N in Figure 1.

Note that when $k=N$, that is a random topic is returned on each website, this measure is 0, as there is no cross-site information leaked. As k decreases, this number grows. When $k=1$ (i.e. always returning the top topic), the cross-site information leakage is $\log N$ bits. These two examples give a sense of the trade-off used in setting parameter k.

A natural question is what is the expected information leakage for the Topics API in the wild. In Section 6 we measure the information leakage for the topics based on real browsing behavior.

Random Topics In the previous section we mostly ignored the effects of p, the probability that a uniformly random topic is returned by the API.

By setting a non-zero p, we ensure that each topic has some minimum support, of roughly $p/N \sim 0.014\%$ of the population for $N=349$ and $p=0.05$. This minimum support gives some protection via k-anonymity of topics, and also helps to move away from the worst case information leakage examples.

5 Longitudinal Analysis

We have considered the amount of cross-site information an API can release on a single call. Due to the persistence of first party ids, an attacker could associate a sequence of values returned by the API for a user on different epochs, which we refer to as a longitudinal profile. This longitudinal profile provides a more powerful way of discovering the identity of a user on another website. Formally, let R be the number of *rounds* an attacker gets to observe the output of the Topics API for a user. Let T_i^r denote the topic observed on website i at round r, $\overline{T}_i = (T_i^1,\ldots,T_i^R)$ and $S_r := S_r(U,k)$ denote the top-k set of interests of a random user at round r. As before, we can show that the cross-site information leakage of the API using a sequence of topics is given by the following mutual information

$$
I(\overline{T}_1;\overline{T_2}). \tag{3}
$$

Notice that when there is no dependency between the top topics associated with a user across rounds we can easily express (3) as:

$$
I(\overline{T}_1;\overline{T}_2) = \sum_{r=1}^{R} I(T_1^r;T_2^r).
$$

Nevertheless, we expect some correlation of top topics across time. The following theorem describes how these correlations affect the mutual information. (We give the proof in Appendix 10.)

Theorem 3. *Let $T_1^{1:r}$ be the sequence of random variables $T_1^{1:r} = T_1^1,\ldots,T_1^r$. We define similarly $T_2^{1:r}$. Using this notation we prove the following bound:*

$$
I(\overline{T_1};\overline{T_2}) \leq \sum_{r=1}^{R} I(T_1^r;T_2^r) + \sum_{r=2}^{R} I\Big((T_1^r,T_2^r)\ ;\ (T_1^{1:r-1},T_2^{1:r-1})\Big).
$$

Let us review the implications of this statement. The theorem shows that the total mutual information of the sequences observed after R rounds is bounded by the sum of two terms. The first is the sum of the per-instance mutual information of observing the two topics, at a given time.

The second term depends on the mutual information of two distributions: (1) the distribution of the pair of topics observed at time r; and (2) the distribution of the pairs of topics observed up to time $r-1$. In the general case, the second term shows that, whenever the correlation with past history is bounded, the additional leakage over time is small. As before, we now try to understand the worst case information leakage across rounds.

Theorem 4 (Worst-case Analysis)**.** *Let $R > 0$ denote the number of topics released. Let $\mathcal{P}$ denote the space of probability measures over collections of top-k sets $(S_1, \ldots, S_R)$ The mutual information between $\overline{T_1}$ and $\overline{T_2}$ can be bounded as:*

$$\max_{P \in \mathcal{P}} I(\overline{T_1}; \overline{T_2}) \leq R \log \frac{N}{k}$$

That is, the amount of information leakage grows linearly with the number of rounds and is at most the per-round maximum information leakage. That is, after R rounds we could potentially leak approximately $6.12R$ bits.

Proof. By the chain rule of mutual information we have

$$I(\overline{T_1}; \overline{T_2}) = \sum_{r=1}^{R} I(\overline{T_1}; T_2^r | T_2^1, \ldots, T_2^{r-1})$$

We now proceed to bound each element in the above summation. By definition of conditional mutual information we have:

$$I(\overline{T_1}; T_2^r | T_2^1, \ldots, T_2^{r-1}) = H(T_2^r | T_2^1, \ldots, T_2^{r-1}) - H(T_2^r | T_2^1, \ldots, T_2^{r-1}, \overline{T}_1). \tag{4}$$

The first term in the above equation is bounded by $\log N$ since T_2^r can take at most N values. We proceed to lower bound the second term. Using the fact that entropy decreases through conditioning we have

$$H(T_2^r | T_2^1, \ldots, T_2^{r-1}, \overline{T}_1) \geq H(T_2^r | T_2^1, \ldots, T_2^{r-1}, \overline{T}_1, S_r) = H(T_2^r | S_r),$$

where we used the fact that the distribution of T_2^r is independent of any other topic observation given the top set S_r. Finally, since T_2^r is a uniform draw from S_r and $|S_r| = k$ we have $H(T_2^r | S_r) = \log k$. Putting it all together yields the desired result. □

Given the previous results we have established that, under the parameters of the model evaluated, the leakage is between $0.97R$ in the optimistic and no correlation scenario and $6.12R$ bits in the worst case.

6 Empirical evaluation of Topics Mutual Information

The examples described in previous sections show that there is a very big range between the best and worst case scenarios for mutual information leakage. It is natural to ask where in that range we expect the real world distribution to occur. To measure this we evaluated the Topics API empirically.

Dataset. The dataset used for this simulation is derived from synced Chrome browsing history. All users represented in the study have opted in to syncing their Chrome history and using Chrome history for cross-product personalization (supplemental Web & App activity). On top of following Google's highest data security and privacy standards, all unique user identifiers were removed before any processing. We simulated the Topics API on these browsing traces and removed users who did not have at least 1 topic associated to them on each of the two week study period and only aggregated results were used for further analysis.

Number of weeks observed (R)	Information leakage		
	No correlation case	Worst case	Observed in Chrome data
1	0.97 bits	6.12 bits	1.01 bits
2	1.94 bits	12.24 bits	2.12 bits

Table 1: Comparison of cross-site information leakage for 1 and 2 weeks of history for observed topics on two different sites.

Taxonomy. We classified websites into topics using a server side version of Chrome's site classifier and used the taxonomy published by Chrome [2].

Time. We simulated the Topics API for $R = 1$ and $R = 2$ weeks.

Simulation. For each user we extracted the sets S_1, S_2 of top 5 topics for each week (sorted by frequency, breaking ties arbitrarily). If a user did not have 5 topics associated with them we padded the set S with random topics from the taxonomy. For each user, we used the top sets to generate a sample of random variables T_i^j for $i = 1, 2$ and $j = 1, 2$. We used this data to generate the joint empirical probabilities of topics across two websites and across two weeks. In this simulation we set the random topic probability, p to zero to match the setup of the previous sections.

Results. We first calculated the *per instance mutual* information of topics across two sites and found it to be 1.01 bits. This value is remarkably close to the uniform scenario discussed in Example 2. However, it is worth noting that the distribution of topics is not uniform since the entropy of the random variable T_1 is 7.09 bits as opposed to 8.44 bits of a uniform random variable. That is, there exists some correlation between the top topics associated with a user, yet it is significantly less than the worst case scenario.

We also analyzed the mutual information of observing two topics (one for each week) and observed a mutual information of 2.12 bits. While this does not match the optimistic scenario of having topics completely independent across time — which would yield $2(1.01) = 2.02$ bits — the result shows that there is limited dependency of topics across two rounds.

7 Limitations

Both our analytical and empirical analysis shed light on the exact privacy leakage of the Topics API; however there are limitations to the conclusions we can draw.

From an analytical standpoint, we made some simplifying assumptions in the behavior of the API. For instance, we did not take into account the 'filtering' of topics that happens when the top topic has not been seen by an ad-tech. Second, we considered the information leaked to each caller separately. If multiple callers were to collude, this increases the overall privacy leakage. For instance, combining the two, if the API is called by m attackers and they all share the value observed, the output of the Topics API is not only a topic t but a binary string that encodes whether each attacker had previously called the API on a page of topic t. This binary string may provide more information about a user's identity than a single topic. Full characterization of the information leakage due to this type of attack as well as the different collusion models that could take advantage of the API are part of our ongoing research.

Empirically, our dataset does not represent a uniform sample of browsing behavior eligible for the Topics API. This is due to both a non-uniform sample of users, as well as the lack of visibility into the potential callers of the API. In addition, we would like to understand the information leakage for $R > 2$. Here we run into issues of scalability and data sparsity. Notice that the state space needed to evaluate the mutual information exactly grows exponentially with R, thus it is harder to estimate the value of each element of the space (sparsity), as well as harder to compute the full mutual information (scalability). Here too, better estimation remains an open area of research.

Finally, there are details of how the API will be used in practice. For instance, our analysis considers a single topic revealed by the Topics API, whereas the API reveals the past three historical topics. Thus the initial information revelation can be analyzed using $R = 3$, but the steady-state rate is at one topic per week, as analyzed.

8 Conclusion

In this note we formalize an attack model on the Topics API. We also introduced the mutual information as a natural measure of re-identification risk. This measure allowed us to understand how the distribution of topics across users, as well as the parameters of the API affect the information leakage. Our analysis presents both worst case and optimistic scenarios, and we provide empirical evidence that on average, real world usage is closer to the optimistic scenario than the worst-case one. In this work we focused on the information theoretic aspects of the topics API. That is, even if after accumulating topics across R rounds, an attacker can, in theory, re-identify users across two sites, there may not be a computationally tractable attack to do so. Our future work will study the effect of potential abuses and misuses of the API, which re-identification attacks are viable in practice, as well as, preventative measures for those attacks.

References

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2001.

[2] J. Karlin. Chrome's topic taxonomy v1, 2022.

[3] E. Rescorla and M. Thomson. Technical comments on floc privacy. Available at `http://https://mozilla.github.io/ppa-docs/floc_report.pdf` (2021/06/10).

9 Proof of Theorem 1

Proof. A simple application of Bayes rule shows that (1) is given by:

$$P(\mathsf{i}_1(U) = i|\mathfrak{T}_1 = \mathfrak{t}_1, T_2 = t_2) = \frac{P(T_2 = t_2, \mathsf{i}_1(U) = i|\mathfrak{T}_1 = \mathfrak{t}_1)}{P(T_2 = t_2|\mathfrak{T}_1 = \mathfrak{t}_1)}$$
$$= \frac{P(T_2 = t_2|\mathsf{i}_1(U) = i, \mathfrak{T}_1 = \mathfrak{t}_1)P(\mathsf{i}_1(U) = i|\mathfrak{T}_1 = \mathfrak{t}_1)}{P(T_2 = t_2|\mathfrak{T}_1 = \mathfrak{t}_1)}$$

Notice that both U and T_2 are independent of $\mathfrak{T}_1$. Therefore the above expression is reduced to:

$$\frac{P(T_2 = t_2|\mathsf{i}_1(U) = i, \mathfrak{T}_1 = \mathfrak{t}_1)P(\mathsf{i}_1(U) = i)}{P(T_2 = t_2)}$$

It is not hard to see that T_2 depends on $\mathsf{i}_1(U)$ and $\mathfrak{T}_1$ only through $T_1 = \mathsf{T}(w_1, U)$ and $t_{1i} = \mathsf{T}(w_1, u_i)$ where u_i is the unique user such that $\mathsf{i}_1(u_i) = i$.

$$\frac{P(T_2 = t_2|T_1 = t_{1i})P(\mathsf{i}_1(U) = i)}{P(T_2 = t_2)}$$

We thus have that the KL-divergence between the two distributions we consider is given by:

$$\sum_i \frac{P(T_2 = t_2|T_1 = t_{1i})P(\mathsf{i}_1(U) = i)}{P(T_2 = t_2)} \log\left(\frac{P(T_2 = t_2|T_1 = t_{1i})}{P(T_2 = t_2)}\right)$$
$$= -\sum_i \frac{P(T_2 = t_2|T_1 = t_{1i})P(\mathsf{i}_1(U) = i)}{P(T_2 = t_2)} \log P(T_2 = t_2)$$
$$+ \sum_i \frac{P(T_2 = t_2|T_1 = t_{1i})P(\mathsf{i}_1(U) = i)}{P(T_2 = t_2)} \log\left(P(T_2 = t_2|T_1 = t_{1i})\right)$$
$$= I_1 + I_2.$$

We proceed to simplify I_1

$$-\sum_i \frac{P(T_2 = t_2|T_1 = t_{1i})P(\mathsf{i}_1(U) = i)}{P(T_2 = t_2)} \log P(T_2 = t_2)$$
$$= -\sum_{t_1 \in \mathcal{V}} \sum_{i:\ t_{1i} = t_1} \frac{1}{n} \frac{P(T_2 = t_2|T_1 = t_1)}{P(T_2 = t_2)} \log P(T_2 = t_2)$$
$$= -\sum_{t_1 \in \mathcal{V}} \frac{n_{t_1}}{n} \frac{P(T_2 = t_2|T_1 = t_1)}{P(T_2 = t_2)} \log P(T_2 = t_2),$$

where n_{t_1} denotes the number of users in w_1 such that $\mathsf{T}(w_1, u) = t_1$. Taking expectation of the above expression over the randomness in w_1 yields

$$-\sum_{t_1 \in \mathcal{V}} \mathbb{E}\left[\frac{n_{t_1}}{n}\right] \frac{P(T_2 = t_2|T_1 = t_1)}{P(T_2 = t_2)} \log P(T_2 = t_2)$$
$$= -\sum_{t_1 \in \mathcal{V}} \frac{P(T_2 = t_2|T_1 = t_1)P(T_1 = t_1)}{P(T_2 = t_2)} \log P(T_2 = t_2) =$$
$$= -\log P(T_2 = t_2),$$

where we used the law of total probability to simplify the above sum. Finally, taking expectation over t_2 we have that

$$\mathbb{E}_{t_2, \mathfrak{t}_1}[I_1] = -\sum_{t_2 \in \mathcal{V}} P(T_2 = t_2) \log P(T_2 = t_2) := H(T_2), \tag{5}$$

where $H(T_2)$ denotes the entropy of T_2. Let us now simplify I_2. A simple application of Bayes rule shows that I_2 is given by

$$\sum_i \frac{1}{n} \frac{P(T_1 = t_{1i}, T_2 = t_2)}{P(T_2 = t_2)P(T_1 = t_{1i})} \log\left(\frac{P(T_1 = t_{1i}, T_2 = t_2)}{P(T_1 = t_{1i})}\right) =$$
$$\sum_{t_1 \in \mathcal{V}} \sum_{i:\ t_{1i} = t_1} \frac{1}{n} \frac{P(T_1 = t_{1i}, T_2 = t_2)}{P(T_2 = t_2)P(T_1 = t_{1i})} \log\left(\frac{P(T_1 = t_1, T_2 = t_2)}{P(T_1 = t_{1i})}\right) =$$
$$\sum_{t_1 \in \mathcal{V}} \frac{n_{t_1}}{n} \frac{P(T_1 = t_1, T_2 = t_2)}{P(T_2 = t_2)P(T_1 = t_1)} \log\left(\frac{P(T_1 = t_{1i}, T_2 = t_2)}{P(T_1 = t_1)}\right)$$

As before, taking expectation with respect to t_1 and t_2 yields

$$\sum_{t_1 \in \mathcal{V}} \sum_{t_2 \in \mathcal{V}} P(T_1 = t_1, T_2 = t_2) \log\left(\frac{P(T_1 = t_1, T_2 = t_2)}{P(T_1 = t_1)}\right)$$

Putting the expression for the expectation of I_1 and I_2 together we see that

$$\begin{aligned} I(U, T_2|\mathfrak{T}_1) &= H(T_2) + \sum_{t_1 \in \mathcal{V}} \sum_{t_2 \in \mathcal{V}} P(T_1 = t_1, T_2 = t_2) \log\left(\frac{P(T_1 = t_1, T_2 = t_2)}{P(T_1 = t_1)}\right) \\ &= H(T_2) - H(T_2|T_1) = I(T_1; T_2). \end{aligned}$$

□

10 Proof of Theorem 3

Proof. By definition

$$I(\overline{T_1}; \overline{T_2}) = \sum_{t_1, t_2} \Pr(\overline{T_1} = t_1, \overline{T_2} = t_2) \log\left(\frac{\Pr(\overline{T_1} = t_1, \overline{T_2} = t_2)}{\Pr(\overline{T_1} = t_1)\Pr(\overline{T_2} = t_2)}\right).$$

We focus on $\frac{\Pr(\overline{T_1}=t_1, \overline{T_1}=t_2)}{\Pr(\overline{T_1}=t_1)\Pr(\overline{T_2}=t_2)}$.

$$\frac{\Pr(\overline{T_1} = t_1, \overline{T_1} = t_2)}{\Pr(\overline{T_1} = t_1)\Pr(\overline{T_2} = t_2)} = \frac{\prod_{r=1}^{R} \Pr(T_1^r = t_1^r, T_2^r = t_2^r | t_1^{1:r-1}, t_2^{1:r-1})}{\prod_{r=1}^{R} \Pr(T_1^r = t_1^r | t_1^{1:r-1}) \prod_{r=1}^{R} \Pr(T_2^r = t_2^r | t_2^{1:r-1})},$$

where we use $t_i^{1:r}$ to indicated the subsequence $t_i^1, \ldots, t_i^r$ of the realization t_i of the variables $T_i^{1:r}$, and where $T_i^{1:0}$ indicates the empty sequence.

By some algebraic manipulation[3], we can show that

$$\begin{aligned} &\frac{\Pr(\overline{T_1} = t_1, \overline{T_1} = t_2)}{\Pr(\overline{T_1} = t_1)\Pr(\overline{T_2} = t_2)} = \\ &= \prod_{r=1}^{R} \frac{\Pr(T_1^r = t_1^r, T_2^r = t_2^r)}{\Pr(T_1^r = t_1^r)\Pr(T_2^r = t_2^r)} \cdot \prod_{r=1}^{R} \frac{\Pr(T_1^r = t_1^r, T_2^r = t_2^r | t_1^{1:r-1}, t_2^{1:r-1})}{\Pr(T_1^r = t_1^r, T_2^r = t_2^r)} \cdot \\ &\prod_{r=1}^{R} \frac{\Pr(T_1^r = t_1^r)}{\Pr(T_1^r = t_1^r | t_1^{1:r-1})} \cdot \prod_{r=1}^{R} \frac{\Pr(T_2^r = t_2^r)}{\Pr(T_2^r = t_2^r | t_2^{1:r-1})} \end{aligned}$$

[3]Here, for simplicity, we make the assumption that for any conditioning and any value of the distributions involved, the probability is non-zero, so the fractions are always well-defined. Notice that this is true in the Topics system due to the uniform randomization with probability p.

By plugging in the previous expression we have,

$$I(\overline{T_1};\overline{T_2}) = \sum_{t_1,t_2} \Pr(\overline{T_1}=t_1,\overline{T_2}=t_2)\log\left(\prod_{r=1}^{R}\frac{\Pr(T_1^r=t_1^r,T_2^r=t_2^r)}{\Pr(T_1^r=t_1^r)\Pr(T_2^r=t_2^r)}\right) \tag{6}$$

$$+\sum_{t_1,t_2} \Pr(\overline{T_1}=t_1,\overline{T_2}=t_2)\log\left(\prod_{r=1}^{R}\frac{\Pr(T_1^r=t_1^r,T_2^r=t_2^r|t_1^{1:r-1},t_2^{1:r-1})}{\Pr(T_1^r=t_1^r,T_2^r=t_2^r)}\right) \tag{7}$$

$$-\sum_{t_1,t_2} \Pr(\overline{T_1}=t_1,\overline{T_2}=t_2)\log\left(\prod_{r=1}^{R}\frac{\Pr(T_1^r=t_1^r|t_1^{1:r-1})}{\Pr(T_1^r=t_1^r)}\right) \tag{8}$$

$$-\sum_{t_1,t_2} \Pr(\overline{T_1}=t_1,\overline{T_2}=t_2)\log\left(\prod_{r=1}^{R}\frac{\Pr(T_2^r=t_2^r|t_2^{1:r-1})}{\Pr(T_2^r=t_2^r)}\right) \tag{9}$$

We address each of the four summands at a time.

We now focus on the first summand (6) and show that it can be written as $\sum_{r=1}^{R} I(T_1^r;T_2^r)$.

$$\sum_{t_1,t_2} \Pr(\overline{T_1}=t_1,\overline{T_2}=t_2)\log\left(\prod_{r=1}^{R}\frac{\Pr(T_1^r=t_1^r,T_2^r=t_2^r)}{\Pr(T_1^r=t_1^r)\Pr(T_2^r=t_2^r)}\right) =$$

$$= \sum_{t_1,t_2} \Pr(\overline{T_1}=t_1,\overline{T_2}=t_2)\sum_{r=1}^{R}\log\left(\frac{\Pr(T_1^r=t_1^r,T_2^r=t_2^r)}{\Pr(T_1^r=t_1^r)\Pr(T_2^r=t_2^r)}\right)$$

$$= \sum_{r=1}^{R}\sum_{t_1,t_2} \Pr(\overline{T_1}=t_1,\overline{T_2}=t_2)\log\left(\frac{\Pr(T_1^r=t_1^r,T_2^r=t_2^r)}{\Pr(T_1^r=t_1^r)\Pr(T_2^r=t_2^r)}\right)$$

$$= \sum_{r=1}^{R}\sum_{t_1^r,t_2^r} \Pr(T_1^r=t_1^r,T_2^r=t_2^r)\log\left(\frac{\Pr(T_1^r=t_1^r,T_2^r=t_2^r)}{\Pr(T_1^r=t_1^r)\Pr(T_2^r=t_2^r)}\right)$$

$$= \sum_{r=1}^{R} I(T_1^r;T_2^r)$$

We now focus on (7)

$$\sum_{t_1,t_2} \Pr(\overline{T_1} = t_1, \overline{T_2} = t_2) \log \left(\prod_{r=1}^{R} \frac{\Pr(T_1^r = t_1^r, T_2^r = t_2^r | t_1^{1:r-1}, t_2^{1:r-1})}{\Pr(T_1^r = t_1^r, T_2^r = t_2^r)} \right)$$
$$= \sum_{r=2}^{R} \sum_{t_1,t_2} \Pr(\overline{T_1} = t_1, \overline{T_2} = t_2) \log \left(\frac{\Pr(T_1^r = t_1^r, T_2^r = t_2^r | t_1^{1:r-1}, t_2^{1:r-1})}{\Pr(T_1^r = t_1^r, T_2^r = t_2^r)} \right)$$
$$= \sum_{r=2}^{R} \sum_{t_1^{1:r}, t_2^{1:r}} \Pr(T_1^{1:r} = t_1^{1:r}, T_2^{1:r} = t_2^{1:r}) \log \left(\frac{\Pr(T_1^r = t_1^r, T_2^r = t_2^r | t_1^{1:r-1}, t_2^{1:r-1})}{\Pr(T_1^r = t_1^r, T_2^r = t_2^r)} \right)$$
$$= \sum_{r=2}^{R} \sum_{t_1^{1:r-1}, t_2^{1:r-1}} \Pr(T_1^{1:r-1} = t_1^{1:r-1}, T_2^{1:r-1} = t_2^{1:r-1}) \cdot$$
$$\sum_{t_1^r, t_2^r} \Pr(T_1^r = t_1^r, T_2^r = t_2^r | T_1^{1:r-1} = t_1^{1:r-1}, T_2^{1:r-1} = t_2^{1:r-1})) \log \left(\frac{\Pr(T_1^r = t_1^r, T_2^r = t_2^r | t_1^{1:r-1}, t_2^{1:r-1})}{\Pr(T_1^r = t_1^r, T_2^r = t_2^r)} \right)$$
$$= \sum_{r=2}^{R} \mathbb{E}_{T_1^{1:r-1}, T_2^{1:r-1}} KL(T_1^r, T_2^r \,|\, T_1^{1:r-1}, T_2^{1:r-1} \,||\, T_1^r, T_2^r)$$
$$= I((T_1^r, T_2^r)\ ;\ (T_1^{1:r-1}, T_2^{1:r-1}))$$

We now focus on the negation of (8); the case of (9) follows by symmetry.

$$\sum_{t_1,t_2} \Pr(\overline{T_1} = t_1, \overline{T_2} = t_2) \log \left(\prod_{r=1}^{R} \frac{\Pr(T_1^r = t_1^r | t_1^{1:r-1})}{\Pr(T_1^r = t_1^r)} \right)$$
$$= \sum_{t_1} \Pr(\overline{T_1} = t_1) \log \left(\prod_{r=1}^{R} \frac{\Pr(T_1^r = t_1^r | t_1^{1:r-1})}{\Pr(T_1^r = t_1^r)} \right)$$
$$= \sum_{r=2}^{R} \sum_{t_1} \Pr(\overline{T_1} = t_1) \log \left(\frac{\Pr(T_1^r = t_1^r | t_1^{1:r-1})}{\Pr(T_1^r = t_1^r)} \right)$$
$$= \sum_{r=2}^{R} \sum_{t_1^{1:r-1}} \Pr(T_1^{1:r-1} = t_1^{1:r-1}) \cdot$$
$$\sum_{t_1^r} \Pr(T_1^r = t_1^r | T_1^{1:r-1} = t_1^{1:r-1})) \log \left(\frac{\Pr(T_1^r = t_1^r | t_1^{1:r-1})}{\Pr(T_1^r = t_1^r)} \right)$$
$$= \sum_{r=2}^{R} \mathbb{E}_{T_1^{1:r-1}} KL(T_1^r \,|\, T_1^{1:r-1} \,||\, T_1^r)$$
$$= \sum_{r=2}^{R} I(T_1^r; T_1^{1:r-1})$$

Now, since the mutual information of any pair of probably distributions is non-negative, we have that both (8) and (9) are non-positive.

This completes the proof of the theorem. □