# Online Appendix to: Personalizing Top-k Processing Online in a Peer-to-Peer Social Tagging Network

XIAO BAI, Yahoo Labs Barcelona
RACHID GUERRAOUI, Ecole Polytechnique Fédérale de Lausanne
ANNE-MARIE KERMARREC, INRIA Bretagne-Atlantique

This online appendix contains complementary proofs for Theorem 4.2 and a comparison of analytical and experimental results.

## A. COMPLEMENTARY PROOFS FOR THEOREM 4.2

### A.1. Computation of $P_S^{(c)}$

$P_S^{(c)}$ is defined as the probability of discovering the ideal interest-based view of $S$ neighbors after $c$ gossip cycles, as we have seen in the proof of Theorem 4.2. According to Theorem 4.1, the state of user $u$'s interest-based view after $c$ gossip cycles can be described as $P^{(c)} = M^c P^{(0)}$ with $P^{(0)}$ the initial state of $u$'s interest-based view and $M$ the transition matrix of a gossip. We know from the proof of Theorem 4.2 that $M$ can be approximated by

$$M_a = \begin{bmatrix} p & 0 & 0 & \cdots & 0 & 0 \\ 1-p & p & 0 & \cdots & 0 & 0 \\ 0 & 1-p & p & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 1-p & p & 0 \\ 0 & 0 & 0 & 0 & 1-p & 1 \end{bmatrix},$$

where $p$ is the probability for user $u$ to not discover any new neighbor for her interest-based view after a gossip.


DOI: http://dx.doi.org/10.1145/2602572

Using mathematical induction, we have

$$
\begin{aligned}
P^{(1)} &= M_a P^{(0)} = \left[p \;\; 1-p \;\; 0 \;\; \cdots \;\; 0\right]^T,\\
P^{(2)} &= M_a P^{(1)} = \left[p^2 \;\; 2p(1-p) \;\; (1-p)^2 \;\; 0 \;\; \cdots \;\; 0\right]^T,\\
P^{(3)} &= M_a P^{(2)} = \left[p^3 \;\; 3p^2(1-p) \;\; 3p(1-p)^2 \;\; (1-p)^3 \;\; 0 \;\; \cdots \;\; 0\right]^T,\\
&\vdots \quad \vdots\\
P^{(S)} &= M_a P^{(S-1)} = \left[\tbinom{S}{0}p^S \;\; \tbinom{S}{1}p^{S-1}(1-p)^1 \;\; \cdots \;\; \tbinom{S}{S-1}p^1(1-p)^{S-1} \;\; \tbinom{S}{S}(1-p)^S\right]^T,\\
P^{(S+1)} &= M_a P^{(S)} = \left[\tbinom{S+1}{0}p^{S+1} \;\; \tbinom{S+1}{1}p^S(1-p)^1 \;\; \cdots \;\; \tbinom{S+1}{S}p^1(1-p)^S \;\; (1+Sp)(1-p)^S\right]^T,\\
P^{(S+2)} &= M_a P^{(S+1)}\\
&= \left[\tbinom{S+2}{0}p^{S+2} \;\; \tbinom{S+2}{1}p^{S+1}(1-p)^1 \;\; \cdots \;\; \tbinom{S+2}{S+1}p^1(1-p)^{S+1} \;\; (1+Sp+\tfrac{S(S-1)}{2}p^2)(1-p)^S\right]^T,\\
&\vdots \quad \vdots\\
P^{(c)} &= M_a P^{(c-1)}\\
&= \left[\tbinom{c}{0}p^c \;\; \tbinom{c}{1}p^{c-1}(1-p)^1 \;\; \cdots \;\; \tbinom{c}{c-S+1}p^{S-1}(1-p)^{c-S+1} \;\; \sum_{x=0}^{c-S}\tbinom{x+S-1}{S-1}p^x(1-p)^S\right]^T.
\end{aligned}
$$

Therefore, the probability $P_S^{(c)}$ of discovering the ideal interest-based view of $S$ neighbors after $c$ gossip cycles can be written as

$$
P_S^{(c)} = \begin{cases} 0 & c < S\\ (1-p)^S & c = S,\\ (1-p)^S \sum_{z=0}^{c-S}\binom{z+S-1}{S-1}p^z, & c > S. \end{cases}
$$

### A.2. Approximation of $P_S^{(c)}$ for $c > S$

We know from the proof of Theorem 4.2 that the probability $P_S^{(c)}$ of discovering the ideal interest-based view of $S$ neighbors after $c$ gossip cycles is $(1-p)^S \sum_{z=0}^{c-S}\binom{z+S-1}{S-1}p^z$ for $c > S$, where $p$ is the probability for user $u$ to not discover any new neighbor for her interest-based view after a gossip.

To simplify this formula, let $f(k,S) = \sum_{z=0}^{k}\binom{z+S-1}{S-1}p^z$, we have

$$
\begin{aligned}
f(k,S) &= \binom{S-1}{S-1} + \binom{S}{S-1}p + \binom{S+1}{S-1}p^2 + \ldots + \binom{k+S-1}{S-1}p^k\\
&= \binom{S-1}{S-1} + (\binom{S}{S-1} + \binom{S}{S})p + (\binom{S+1}{S-1} + \binom{S+1}{S})p^2 + \ldots +\\
&\quad (\binom{k+S-1}{S-1} + \binom{k+S-1}{S})p^k - (\binom{S}{S}p + \binom{S+1}{S}p^2 + \ldots + \binom{k+S-1}{S}p^k).
\end{aligned}
$$

With the property of binomial coefficients, $\binom{n}{m} = \binom{n-1}{m-1} + \binom{n-1}{m}$ for $0 < m < n$, we have

$$\begin{aligned} f(k,S) &= \binom{S}{S} + \binom{S+1}{S}p + \binom{S+2}{S}p^2 + ... + \binom{k+S}{S}p^k \\ &\quad -p(\binom{S}{S} + \binom{S+1}{S}p + ... + \binom{k+S-1}{S}p^{k-1}) \\ &= f(k,S+1) - pf(k-1,S+1) \\ &= f(k,S+1) - pf(k,S+1) + p\binom{k+S}{S}p^k. \end{aligned}$$

We obtain the recurrence formula

$$f(k,S+1) = \frac{1}{1-p}f(k,S) - \frac{1}{1-p}\binom{k+S}{S}p^{k+1}.$$

Therefore,

$$\begin{aligned} f(k,S) &= \frac{1}{1-p}f(k,S-1) - \frac{1}{1-p}\binom{k+S-1}{S-1}p^{k+1} \\ &= \frac{1}{1-p}(\frac{1}{1-p}f(k,S-2) - \frac{1}{1-p}\binom{k+S-2}{S-2}p^{k+1}) - \frac{1}{1-p}\binom{k+S-1}{S-1}p^{k+1} \\ &= \cdots \\ &= \frac{1}{(1-p)^{S-1}}f(k,1) - p^{k+1}\sum_{l=1}^{S-1}\frac{1}{(1-p)^l}\binom{k+S-l}{S-l}. \end{aligned}$$

Let $g(l) = \frac{1}{(1-p)^l}\binom{k+S-l}{S-l}$. Since $(\frac{n}{m})^m \le \binom{n}{m} \le (\frac{en}{m})^m$, we have

$$g(l) \le \frac{1}{(1-p)^l}(\frac{e(k+S-l)}{S-l})^{S-l}.$$

Letting the right side of this inequality be $h(l)$, we can prove that $h(l)$ is monotonically decreasing for $1 \le l \le S-1$, since its derivative is negative for these values of $l$. Hence, we obtain a lower bound of $f(k,S)$, that is,

$$f(k,S) \ge \frac{1}{(1-p)^{S-1}}f(k,1) - p^{k+1}\sum_{l=1}^{S-1}h(l) \ge \frac{1}{(1-p)^{S-1}}f(k,1) - p^{k+1}(S-1)h(1).$$

Since

$$f(k,1) = 1 + p + p^2 + ... + p^k = \frac{1-p^{k+1}}{1-p},$$

$$h(1) = \frac{1}{1-p}(\frac{e(k+S-1)}{S-1})^{S-1},$$

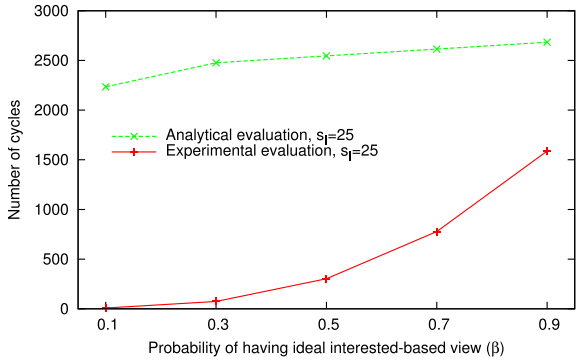


Fig. 15. Comparison of analytical and experimental convergence performance of $P^2TK^2$.

we obtain

$$\begin{aligned} P_S^{(c)} &= (1-p)^S f(c-S,S) \\ &\geq (1-p)^S \left(\frac{1}{(1-p)^{S-1}} \frac{1-p^{c-S+1}}{1-p} - p^{c-S+1}(S-1)\frac{1}{1-p}\left(\frac{e(c-1)}{S-1}\right)^{S-1}\right) \\ &= 1 - p^{c-S+1} - (1-p)^{S-1} p^{c-S+1} \frac{e^{S-1}(c-1)^{S-1}}{(S-1)^{S-2}}. \end{aligned}$$

Therefore, for $c > S$, the probability of discovering the ideal interest-based view of $S$ neighbors after $c$ gossip cycles $P_S^{(c)}$ can be approximated by its lower bound

$$1 - p^{c-S+1} - (1-p)^{S-1} p^{c-S+1} \frac{e^{S-1}(c-1)^{S-1}}{(S-1)^{S-2}}.$$

## B. COMPARISON OF ANALYTICAL AND EXPERIMENTAL RESULTS

To illustrate how the analytical evaluation of $P^2TK^2$ reflects its experimental evaluation, we compare their convergence performance in terms of building interest-based views in this section. Specifically, in the analytical evaluation, given a probability $\beta$, we solve the equation in Theorem 4.2 to obtain the number of cycles that are necessary for a user to discover her ideal interest-based view of $S$ neighbors in a system of $N$ users. This is equivalent to the number of cycles that are necessary for a fraction $\beta$ of $N$ users in the system to discover their ideal interest-based views of $S$ neighbors in the experimental evaluation. In fact, based on the assumption that the gossip behaviors of users are independent Markov processes, this experiment can be considered as repeating the process of building interest-based view for a user $N$ times in parallel. Therefore, the number of users having their ideal interest-based view at a given cycle divided by the total number of users gives the probability of a user having her ideal interest-based view at that cycle.

Figure 15 compares the analytical and experimental results for different values of $\beta$ on the *CiteULike* dataset (Section 5.1). As expected, we observe that the estimated number of cycles to discover the ideal interest-based view is much higher than the measured number of cycles[1] for a given probability $\beta$. This is because we assume random

[1]Note that this measure is different from the number of cycles in Figure 15(a). This is because even if all users have discovered 90% of their ideal interest-based views at a cycle, the number of users who have discovered their ideal interest-based view may still be 0. This measure is thus more demanding.

gossip in the analytical model, while in the experiment, users gossip with neighbors in their interest-based views. This allows them to discover other neighbors with similar interests more easily by leveraging the fact that the friends' friends are likely to be friends, especially in the early stage of the experiment. Yet, as shown in Figure 15, the analytical model provides a theoretical upper bound for the convergence property of $P^2TK^2$. Moreover, the bound is tighter with higher values of $\beta$, which is desirable for the effectiveness of $P^2TK^2$.

We ignore the comparison for the performance of query processing given the similarity of its process to that of building the interest-based views. Similarly, the analytical evaluation would provide a loose upper bound of the necessary number of cycles to resolve user queries.