Supplementary Material for “Generative Models for Item Adoptions Using Social Correlation”

Freddy Chong Tat Chua, Hady W. Lauw, Ee-Peng Lim

Appendix A
Derivation of the E-Steps and M-Steps for Unified Generative Model

Suppose we have Θ the users latent factor distributions and Φ the latent factors item distribution. Then the likelihood of E is given by,

$$
\begin{aligned}
P(E|\Theta, \Phi, C, F) &= \prod_{u \in U} \prod_{v \in V_u} P(e_{v,u}|\Theta, \Phi, C, F) \\
&= \prod_{u \in U} \prod_{v \in V_u} \sum_{z \in Z} \sum_{x \in F_u} \Big[P(e_{v,u}|z_{v,u} = z, \Phi) \\
& P(z_{v,u} = z|x_{v,u} = x, \Theta) P(x_{v,u} = x|C_u, F_u) \Big]
\end{aligned}
$$

Then expressing in logarithm form,

$$
\begin{aligned}
\log P(E|\Theta, \Phi, C) &= \sum_{u \in U} \sum_{v \in V_u} \log \Big[\sum_{z \in Z} \sum_{x \in F_u} P(e_{v,u}|z_{v,u} \\
&= z, \Phi) P(z_{v,u} = z|x_{v,u} = x, \Theta) P(x_{v,u} = x|C_u, F_u) \Big]
\end{aligned}
$$

Find the E Step for $z_{v,u}$ assuming that we do not have $x_{v,u}$,

$$
\begin{aligned}
& P(z_{v,u} = z|e_{v,u}, \Theta, \Phi, C, F) \\
&= \frac{\sum_{x \in F_u} P(e_{v,u}, z, x_{v,u} = x|\Theta, \Phi, C, F)}{\sum_{z' \in Z} \sum_{x' \in F_u} P(e_{v,u}, z', x_{v,u} = x'|\Theta, \Phi, C, F)} \\
&\propto \sum_{x \in F_u} P(e_{v,u}|z, \Phi) P(z|x, \Theta) P(x|C_u, F_u) \\
&= g(u, z, v)
\end{aligned}
$$

Then find the E Step for $x_{v,u}$ assuming that we do not have $z_{v,u}$,

$$
\begin{aligned}
& P(x_{v,u} = x|e_{v,u}, \Theta, \Phi, C, F) \\
&= \frac{\sum_{z \in Z} P(e_{v,u}, z_{v,u} = z, x|\Theta, \Phi, C, F)}{\sum_{z' \in Z} \sum_{x' \in F_u} P(e_{v,u}, z_{v,u} = z', x'|\Theta, \Phi, C, F)} \\
&\propto \sum_{z \in Z} P(e_{v,u}|z, \Phi) P(z|x, \Theta) P(x|C_u, F_u) \\
&= h(u, x, v)
\end{aligned}
$$

In the M Step of EM algorithm, take partial derivative of the log likelihood with respect to Θ, Φ and C,

$$
\log P(E|\Theta, \Phi, C) = \sum_{u \in U} \sum_{v \in V_u} \log \left(\sum_{z \in Z} \sum_{u' \in U} \phi_{z,v} \theta_{u',z} c_{u,u'} \right)
$$

Given that $\sum_{u' \in U} c_{u,u'} = 1$, $\sum_{z \in Z} \theta_{u,z} = 1$ and $\sum_{v \in V_u} \phi_{z,v} = 1$ are constraints, we may optimize for the above using the following Lagrange constraint,

$$
\begin{aligned}
& \mathcal{L}(\Theta, \Phi, C, F, \lambda) = \log P(E|\Theta, \Phi, C, F) \\
& - \sum_{u \in U} \left[\lambda_u \left(\sum_{x \in F_u} c_{u,x} - 1 \right) + \gamma_u \left(\sum_{z \in Z} \theta_{u,z} - 1 \right) \right] \\
& - \sum_{z \in Z} \delta_z \left(\sum_{v \in V_u} \phi_{z,v} - 1 \right)
\end{aligned}
$$

Suppose we differentiate $\mathcal{L}(\Theta, \Phi, C, F, \lambda)$ with respect to $c_{u,x}$, $\theta_{x,z}$ and $\phi_{z,v}$:

$$
\begin{aligned}
& \frac{d}{d\ c_{u,x}} \mathcal{L}(\Theta, \Phi, C, \lambda) \\
&= \sum_{v \in V_u} \frac{\sum_{z \in Z} \phi_{z,v} \theta_{x,z}}{\sum_{z' \in Z} \sum_{x' \in F_u} \phi_{z',v} \theta_{x',z'} c_{u,x'}} - \lambda_u \\
& \frac{d}{d\ \theta_{u,z}} \mathcal{L}(\Theta, \Phi, C, \lambda) \\
&= \sum_{v \in V_u} \frac{\phi_{z,v} c_{u,x}}{\sum_{z' \in Z} \sum_{x' \in F_u} \phi_{z',v} \theta_{x',z'} c_{u,x'}} - \gamma_u \\
& \frac{d}{d\ \phi_{z,v}} \mathcal{L}(\Theta, \Phi, C, \lambda)
\end{aligned}
$$

- *Freddy Chong Tat Chua is a PhD student at the School of Information Systems, Singapore Management University, Singapore. E-mail: freddy.chua.2009@smu.edu.sg*
- *Hady W. Lauw is an Assistant Professor in the School of Information Systems, Singapore Management University, Singapore. E-mail: hadywlauw@smu.edu.sg*
- *Ee-Peng Lim is a Professor in the School of Information Systems, Singapore Management University, Singapore. E-mail: eplim@smu.edu.sg*

$$= \sum_{u \in U} \frac{\sum_{x \in F_u} \theta_{x,z} c_{u,x}}{\sum_{z' \in Z} \sum_{x' \in F_u} \phi_{z',v} \theta_{x',z'} c_{u,x'}} - \delta_z$$

Then find the $c_{u,x}$, $\theta_{u,z}$ and $\phi_{z,v}$ which gives zero gradient for $\mathcal{L}(C, \lambda)$. To summarize, the E Steps are

$$f(u,v,z) = \frac{\phi_{z,v} \theta_{u,z} c_{u,u}}{\sum_{z' \in Z} \sum_{x' \in F_u} \phi_{z',v} \theta_{x',z'} c_{u,x'}} \quad (1)$$

$$g(u,v,z) = \frac{\sum_{x \in F_u} \phi_{z,v} \theta_{x,z} c_{u,x}}{\sum_{z' \in Z} \sum_{x' \in F_u} \phi_{z',v} \theta_{x',z'} c_{u,x'}} \quad (2)$$

$$h(u,v,x) = \frac{\sum_{z \in Z} \phi_{z,v} \theta_{x,z} c_{u,x}}{\sum_{z' \in Z} \sum_{x' \in F_u} \phi_{z',v} \theta_{x',z'} c_{u,x'}} \quad (3)$$

The M Steps are,

$$\theta_{u,z} = \frac{1}{\gamma_u} \sum_{v \in V_u} f(u,v,z) \quad (4)$$

$$\phi_{z,v} = \frac{1}{\delta_z} \sum_{u \in U} g(u,v,z) \quad (5)$$

$$c_{u,x} = \frac{1}{\lambda_u} \sum_{v \in V_u} h(u,v,x) \quad (6)$$

Appendix B Topic Analysis

Here, we evaluate the effectiveness of LDA in deriving the latent factors or topics. If LDA has learned the latent factors or topics well, each topic would correspond to a cluster of related items. For ease of illustration, we only show three topics each for LiveJournal and Epinions. For each topic, we identify the top items with the highest latent factor values for that topic.

Table 5 shows a sample of the top communities in each topic for the LiveJournal data set. The names of communities in LiveJournal draw from a wide variety of languages with Russian being a dominant language as seen by the prefix *ru_* in the communities name. *Topic L1* shows preference for East Asian culture. "jpop" is a synonym for Japanese Pop Music, "kpop" for Korean Pop Music, "jdramas" for Japanese Drama, "anime" and "manga" are terms for Japanese cartoons. *Topic L2* is of Information Technology subjects and *Topic L3* shows art and design. Table 6 shows a sample of the top movie titles in each topic for the Epinions data set. The movies in each topic tend to be similar in terms of their genres. For instance, movies in *Topic E1* such as the Spider-Man and Lord of the Rings series are action movies. Movies in *Topic E2* are dramas such as Erin Brockovich and Fight Club. Movies in *Topic E3* seem to be comedies. Intuitively, these three topics also correspond to the three most popular genres in the data set: action, drama, and comedy.

TABLE 5
Example Top Communities for Each Topic in LiveJournal

Topic L1	*Topic L2*	*Topic L3*
free_manga	ru_webdev	ru_designer
anime_downloads	ru_linux	ru_photoshop
jdramas	ru_sysadmins	design_books
jpop_uploads	ru_software	ru_illustrators
kpop_uploads	ru_programming	ru_vector

TABLE 6
Example Top Movie Titles for Each Topic in Epinions

Topic E1	*Topic E2*	*Topic E3*
Spider-Man	Erin Brockovich	Shrek
Spider-Man 2	Fight Club	Charlie's Angels
Batman Begins	American Psycho	What Women Want
Lord of the Rings: The Two Towers	Magnolia	Meet the Parents
Lord of the Rings: The Return of the King	American Beauty	Miss Congeniality

Appendix C Distribution of Social Correlation

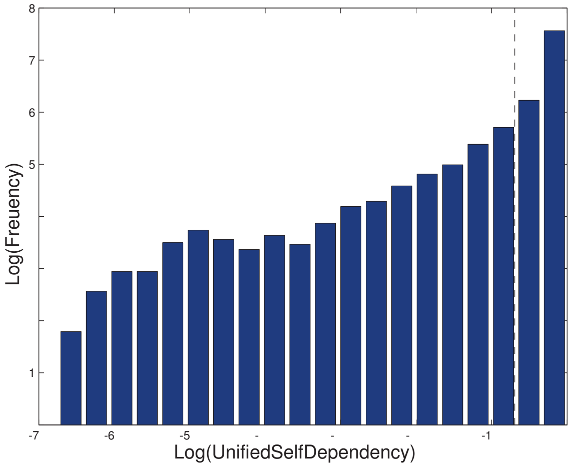

Fig. 17. LiveJournal: Histogram of Self Dependency

Figures 17 and 18 show the histogram of self-dependency values. The x-axis indicates the self-dependency values in logarithm scale and y-axis indicates the number of users who fall into the respective bins. The dotted black line parallel to the y-axis represents the logarithm value of 0.5. We define users having self dependency value less than 0.5 as followers (left of the dotted line), because they depend more on others in aggregate than in themselves. With this definition, 35% of users in LiveJournal and 29% of users in Epinions are followers. These significant percentages indicate that a sizable portion of the population do depend on others in their item adoptions, which validate our proposed approach of not relying on self preferences alone.

On the other hand, since the majority of users are non-followers, many social links between the users

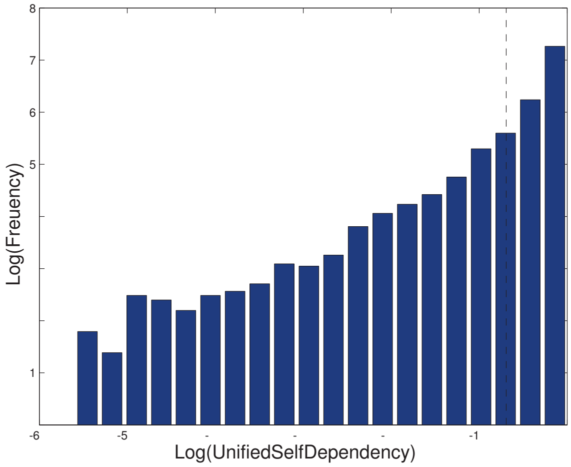

Fig. 18. Epinions: Histogram of Self Dependency

have very low social correlation values. In other words, a user may choose to follow another user but many of such follow relationships do not share common interests or result in item adoptions for the following user. This may imply that while the observed social network is sparse, the actual underlying dependency network between users is sparser.

Appendix D
Theoretical Performance of Random

Given that there are M items for the Random prediction model to select from and v out of M items are Actual Positive. That is, a random user has these v items in the testing set and we want to test how well Random method recovers these v items. Then given that we select the top k items returned by the Random method such that $k \leq M$. What is the probability that there are t correctly chosen items, given that $t \leq v$?

Since AUC of Precision & Recall (AUC-PR) Curve for Random depends on the precision ($PREC$) and recall (REC) for each k, we should find the expected precision $E(PREC|k)$ and expected recall $E(REC|k)$ for each k. Expected values of precision and recall depends on the number of true positives (tp) at k,

$$\begin{aligned} E(PREC|k) &= \frac{E(tp|k)}{k} \\ E(REC|k) &= \frac{E(tp|k)}{v} \\ E(tp|k) &= \sum_{t=1}^{min(k,v)} t \cdot P(tp = t|k) \\ P(tp = t|k) &= \binom{v}{t} \cdot \binom{M-v}{k-t} / \binom{M}{k} \end{aligned}$$

$P(tp = t|k)$ is derived as follows, given that there are v actual positives, the number of possible ways to get t predicted positives, is the combinatorial $\binom{v}{t}$. Then there are $M - v$ actual negatives, to select $k - t$ predicted negatives out of these actual negatives, we have $\binom{M-v}{k-t}$ different combinations of selections. Finally, there are $\binom{M}{k}$ ways of choosing top k randomly from the entire possible set of items.

$P(tp = t|k)$ is in fact a HyperGeometric Distribution. Finally, expected AUC of PR Curve is given by the area under curve of the list of PR values for each k, from 1 to M.

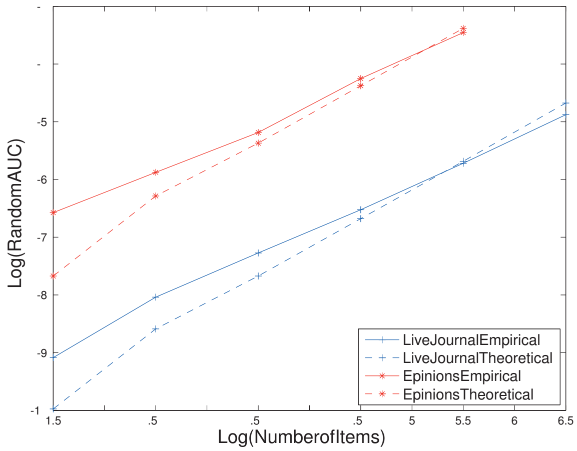

Fig. 19. Log(AUC of Random) vs Log(Number of Items)

Figure 19 shows the theoretical and actual empirical results given by *Random*. The performance of *Random* increases as number of items increases. This explains why our AUC ratio which represents the improvement over *Random* decreases when number of items increases, as shown in Figures 11 and 12. The values of AUC on the y-axis in Figure 19 shows that the AUC values are in the order of e^{-10} to e^{-3}. In comparison, the AUC values obtained by our models as reflected in Figures 5 and 6 are relatively higher than *Random*.