Online Appendix for “Spillover Effects and Freemium Strategy in the Mobile App Market”

Yiting Deng Anja Lambrecht Yongdong Liu

A Identify Freemium Pairs

We use Python to identify freemium pairs in three general steps described below.

Step 1: Match

For each publisher, we generate a group comprising of all apps developed by this publisher.

For each app, we compute its similarity to all other apps in the same groups, using app name as the basis of comparison. For each app pair, we first compute a Jaro-Winkler distance between the full strings of the names. After this, we tokenize app names, leading to two lists of tokens (words) for each app pair. For instance, “Angry Birds” and “Angry Birds Lite” are tokenized into [“angry”, “birds”] and [“angry”, “birds”, “lite”]. We then remove paid/free keywords (e.g., “lite”, “pro”, free”, “premium”) . In the the previous example, the lists become [“angry”, “birds”] and [“angry”, “birds”]. For each pair of words in order (meaning “angry” will only compare to “angry”), we compute a new Jaro-Winkler distance which is added to the previous measure of the full string distance. The aggregated Jaro-Winkler distance is then normalized to a scale of 0-1.

Using this established Jaro-Winkler distance, each app pair gets sent into a series of heuristics using cut-off values to determine the quality of match. Very confident matches (>=0.9) are added as a “NoFlag” match indicating that they are very close matches. Matches that score a distance between 0.9-0.8 are added but with an added “Flag” variable indicating that they are closely related but may contain some false positives. After these two cut-offs, two more samples are added at 0.8-0.75 and 0.75-0.8 respectively. These last two cut-offs add an “OtherSample” variable to the match.

Step 2: Remove False Matches

We remove the following types of false matches:

- We remove matches between an iPad app and an iPhone app using a set of iPad specific keywords (e.g. “Angry Birds” and “Angry Birds iPad” or “Angry Birds HD”).
- Sequels such as “angry birds” and “angry birds 2” would have a very high similarity score but are not a true freemium pair. We remove sequels by comparing suffix numbers of app names. If app1 has a “2” at the end whilst app2 has nothing, app1 is considered a sequel to app2 or vice versa. Sequels up to the number 4 are considered.
- If a free (paid) app matched with another free (paid) app respectively (i.e. paid-paid or free-free matches only), this pair is marked for removal (e.g. “angry birds lite” and “angry birds free”).

Step 3: Manual Check

We order all the identified freemium pairs so that app1 is the paid version and app2 is the free version. We then manually check to ensure that they are indeed a freemium pair.

B Summary Statistics

Table A.1 reports summary statistics for the paid version of freemium apps in the panel setting. For each variable x, we compute the overall, within and between variance as follows.

Denote the value of variable x for app i on day t as x_{it}, $i = 1,...,N$; $t = 1,...,T$. Denote the mean value of x across apps and days as $\bar{x}$, and the mean value of x for app i as $\bar{x}_i$. The overall variance of x is computed as $\frac{\sum_i \sum_t (x_{it} - \bar{x})^2}{NT-1}$. The between variance of x is computed as $\frac{\sum_i (\bar{x}_i - \bar{x})^2}{N-1}$. The within variance of x is computed as $\frac{\sum_i \sum_t (x_{it} - \bar{x}_i + \bar{x})^2}{NT-1}$.

Table A.1: Summary Statistics on Freemium Apps (with Both Within- and Between- Variations)

Variable	Mean	Std. Dev.			Min	Max
		Overall	Between apps	Within app		
Incremental daily No. of ratings	0.54	9.44	4.68	8.20	0	1316
ln (incremental daily No. of ratings+1)	0.07	0.40	0.28	0.28	0	7.2
Price	1.33	1.51	1.13	1.01	0	100
If rated	0.30	0.46	0.44	0.15	0	1
No. ratings	105.18	1209.93	1243.15	39.35	0	29717
ln (No. ratings+1)	1.02	1.78	1.73	0.41	0	10.3
Average star rating	4.26	0.69	0.68	0.06	1.2	5
Ranked on top 10	0.00	0.03	0.01	0.03	0	1
Ranked on top 11-20	0.00	0.02	0.01	0.02	0	1
Ranked on top 21-50	0.00	0.04	0.03	0.03	0	1
Ranked on top 51-100	0.00	0.04	0.02	0.03	0	1
Ranked on top 101-	0.05	0.21	0.17	0.14	0	1
Age (Days)	147.96	214.14	213.92	8.32	1	1694

Table A.2 shows the descriptive statistics on the propensity score matching results, which indicate that the freemium and control apps are not significantly different in key variables.

Table A.2: Descriptive Statistics on the Matching Results

	Means Treated (s.e.)	Means Control (s.e.)	p-value of t-test
Propensity Score	0.0007	0.0007	0.9995
Ave. Price of Week	1.34(1.38)	1.32(1.54)	0.7714
Ln(No of 1-Star Ratings)	0.33(0.95)	0.34(0.98)	0.5297
Ln(No of 2-Star Ratings)	0.25(0.82)	0.26(0.82)	0.6010
Ln(No of 3-Star Ratings)	0.31(0.94)	0.31(0.94)	0.9827
Ln(No of 4-Star Ratings)	0.41(1.09)	0.39(1.1)	0.4985
Ln(No of 5-Star Ratings)	0.77(1.54)	0.7(1.55)	0.1249
Ranked on top 10	0.05%(1.35%)	0.06%(1.83%)	0.8225
Ranked on top 11-20	0.02%(0.39%)	0.05%(1.46%)	0.4372
Ranked on top 21-50	0.15%(2.82%)	0.12%(2.47%)	0.6631
Ranked on top 51-100	0.12%(1.94%)	0.22%(3.35%)	0.2499
Ranked on top 101-	5.00%(18.15%)	5.27%(19.59%)	0.5903
Genre ID	Exact match		
Whether Updated in the 2-week window before free version introduction	Exact match		
Age	Difference within 6 months		

C Analyses on the Effect of Sampling

Effect of $\underline{v}_p$ on $E(U_s)$:

$$\begin{aligned}
\frac{dE(U_s)}{d\underline{v}_p} &= \frac{-2\theta^2\bar{v}_f 2\theta\left(\bar{v}_p-\underline{v}_p\right)+2\theta\left[(p+\theta\bar{v}_f)^2-2\theta^2\bar{v}_f\underline{v}_p-2p\theta\bar{v}_p+(\theta\bar{v}_p)^2\right]}{4\theta^2\left(\bar{v}_p-\underline{v}_p\right)^2} \\
&= \frac{-2\theta^2\bar{v}_f\left(\bar{v}_p-\underline{v}_p\right)+(p+\theta\bar{v}_f)^2-2\theta^2\bar{v}_f\underline{v}_p-2p\theta\bar{v}_p+(\theta\bar{v}_p)^2}{2\theta\left(\bar{v}_p-\underline{v}_p\right)^2} \\
&= \frac{(p+\theta\bar{v}_f)^2-2\theta\bar{v}_p(p+\theta\bar{v}_f)+(\theta\bar{v}_p)^2}{2\theta\left(\bar{v}_p-\underline{v}_p\right)^2} \\
&= \frac{(p+\theta\bar{v}_f-\theta\bar{v}_p)^2}{2\theta\left(\bar{v}_p-\underline{v}_p\right)^2}>0
\end{aligned}$$

Therefore, $E(U_p)$ increases with $\underline{v}_p$.

Effect of $\underline{v}_p$ on $E(U_s)-E(U_p)$:

$$\begin{aligned}
E(U_s)-E(U_p) &= v_f\int_{\theta\underline{v}_p}^{p+\theta\bar{v}_f} f(v_p)\,d(v_p)+\int_{p+\theta\bar{v}_f}^{\theta\bar{v}_p} f(v_p)(v_p-p)\,d(v_p)-c-\left[\int_{\theta\underline{v}_p}^{\theta\bar{v}_p} v_p f(v_p)\,d(v_p)-p\right] \\
&= \frac{(p+\theta\bar{v}_f)^2-2\theta^2\bar{v}_f\underline{v}_p-2p\theta\bar{v}_p+(\theta\bar{v}_p)^2}{2\left(\bar{v}_p-\underline{v}_p\right)}-\frac{\theta\left(\underline{v}_p+\bar{v}_p\right)}{2}+p-c \\
&= \frac{(p+\theta\bar{v}_f)^2-2\theta^2\bar{v}_f\underline{v}_p-2p\theta\bar{v}_p-\theta^2\underline{v}_p^2}{2\theta\left(\bar{v}_p-\underline{v}_p\right)}+p-c
\end{aligned}$$

Thus, we have:

$$\begin{aligned}
\frac{dE(U_s-U_p)}{d\underline{v}_p} &= \frac{dE(U_s)}{d\underline{v}_p}-\frac{dE(U_p)}{d\underline{v}_p} \\
&= \frac{(p+\theta\bar{v}_f-\theta\bar{v}_p)^2}{2\theta\left(\bar{v}_p-\underline{v}_p\right)^2}-\frac{\theta}{2} \\
&= \frac{(p+\theta\bar{v}_f-\theta\bar{v}_p)^2-\theta^2\left(\bar{v}_p-\underline{v}_p\right)^2}{2\theta\left(\bar{v}_p-\underline{v}_p\right)^2}
\end{aligned}$$

Note that:

$$
\begin{aligned}
\left(\theta\bar{v}_p - p - \theta\bar{v}_f\right)^2 - \theta^2\left(\bar{v}_p - \underline{v}_p\right)^2 &= \left(p + \theta\bar{v}_f - \theta\bar{v}_p + \theta\left(\bar{v}_p - \underline{v}_p\right)\right)\left(p + \theta\bar{v}_f - \theta\bar{v}_p - \theta\left(\bar{v}_p - \underline{v}_p\right)\right) \\
&= \left(p + \theta\bar{v}_f - \theta\underline{v}_p\right)\left(p + \theta\bar{v}_f - 2\theta\bar{v}_p + \theta\underline{v}_p\right)
\end{aligned}
$$

Because $\underline{v}_p < p$, $p + \theta\bar{v}_f - \theta\underline{v}_p > 0$

Because $\theta\underline{v}_p \leq v_p \leq \theta\bar{v}_p$, $p + \theta\bar{v}_f - 2\theta\bar{v}_p + \theta\underline{v}_p \leq p + \theta\bar{v}_f - 2v_p + v_p = p + \theta\bar{v}_f - v_p = p - \left(v_p - \theta\bar{v}_f\right)$. If the consumer updates to paid version, then $v_p - \theta\bar{v}_f > p$. Thus $p + \theta\bar{v}_f - 2\theta\bar{v}_p + \theta\underline{v}_p < 0$.

Therefore, $\frac{dE(U_s - U_p)}{d\underline{v}_p} < 0$, indicating that $E\left(U_s\right) - E\left(U_p\right)$ decreases with $\underline{v}_p$.

Effect of $\underline{v}_p$ on the sampling effect: Sampling is effective when $E\left(U_s\right) > 0$ and $E\left(U_s\right) > E\left(U_p\right)$. Assume $E\left(U_s\right) = 0$ when $\underline{v}_p = \underline{v}_p^{\star}$; and $E\left(U_s\right) - E\left(U_p\right) = 0$ when $\underline{v}_p = \underline{v}_p^{\star\star}$.

Because $E\left(U_s\right)$ increases with $\underline{v}_p$ but $E\left(U_s\right) - E\left(U_p\right)$ decrease with $\underline{v}_p$, the consumer both samples and also upgrades to the paid version when $\underline{v}_p^{\star} \leq \underline{v}_p \leq \underline{v}_p^{\star\star}$.

For a piece of information on app quality that shifts $\underline{v}_p$, condition $\underline{v}_p^{\star} \leq \underline{v}_p \leq \underline{v}_p^{\star\star}$ is more likely to meet when the signal indicates that app quality is neither too low nor too high, thus the consumer is most likely to sample and to upgrade when this quality signal is moderate.