Explanation of Unique User Estimation

Parameters

n_{max} Maximum number of preference file creation times tied together in normal use

α Probability a user did not run Flash and Java sims, given an entry with just one preferences file

β Fraction of many-preferences cases that are shared ($0 \Rightarrow$ for all JeffCo cases, $1 \Rightarrow$ for all shared cases)

δ Estimated number of users per computer

Entities

Multiple rows in the user table can have the same preferences_file_creation_time or installation_timestamp. They are from either the same computer or group of computers, and can be associated together. We can group these rows into entities (a group of rows that can be linked to each other) by the following equivalence relation:

row A and row Z are equivalent if either of the following hold:

- A and Z either share the same preferences_file_creation_time OR installation_timestamp
- A is equivalent to B, which is equivalent to ... which is equivalent to Y which is equivalent to Z

For each entity, we can calculate the number of preferences times, the number of installation timestamps, the number of user total sessions (sum the greatest values for each pref time), and first and last seen months.

An entity falls into one of three categories:

unlinked (number of pref times $= 1$**)** A preference file creation time is not linked to any other times, so we can't tell whether either (a) the user ran only Flash or Java, or (b) the user ran Flash and Java, but not both from an installation, so there are two entities for this user.

linked ($n_{max} \geq$ **number of pref times** > 1**)** We know they ran both Flash and Java. (Flex will have the same pref times as Flash). Either way, we are not overcounting due to the unlinked condition.

jeffco or shared ($n_{max} <$ **number of pref times)** This entity may be one of the following (and there is no way to tell which, so we estimate this with beta):

- JeffCo case (one computer). prefs are reset each time, but installation timestamp stays the same. Could be multiple users theoretically, however we simplify it to say one user per installation in this case
- Shared case (multiple computers and users). installation is shared across multiple computers, yet each computer has local storage for prefs, so we have the same situation, but we have many users.

Additional Variables

The following are pulled from the entity table:

U Number of unlinked entities

L Number of linked entities

C Number of shared/jeffco entities

P Number of pref times in all shared/jeffco entities

N is the number of unique users

Formula for user count

$$N \approx \delta \left(\frac{U}{\alpha + 1} + L + \frac{\beta}{2} P + (1 - \beta) \, C \right)$$

All terms are scaled by δ, since the terms estimate the number of computers, then we estimate the number of users from that. The first two terms should carry most of the users, the last two are for estimating the shared and jeffco cases respectively.