Yahtzee: An Anonymized Group Level Matching Procedure

Jason J. Jones Robert M. Bond Christopher J. Fariss Jaime E. Settle Adam Kramer
Cameron Marlow James H. Fowler
University of California San Diego & Facebook

Goals

- Protect the privacy of subjects in social science research
- Combine multiple datasets that contain personal information

Solution

- Keep individual level data from being shared between datasets
- Randomly assign individuals to anonymous groups before combining the anonymized information
- Yahtzee!

Application Preview

- We combine data from Facebook and public voter records without linking individual data. Only group level data are shared between the two datasets

The Yahtzee Method: Individuals to Groups

	first name	last name	date of birth	Salt	concatenated value to hash	last 7 hash digits
1	Jason	Jones	11/07/1977	XKCD	JASONJONES19771107XKCD	b815d72
2	Robert	Bond	10/2/1983	XKCD	ROBERTBOND19831021XKCD	3863afe
3	Christopher	Fariss	11/18/1981	XKCD	CHRISTOPHERFARISS19811118XKCD	e0df6f8
4	Jaime	Settle	7/5/1985	XKCD	JAIMESETTLE19850705XKCD	c2e47b1
5	Adam	Kramer	1/24/1981	XKCD	ADAMKRAMER19810124XKCD	947407f
6	Cameron	Marlow	3/28/1977	XKCD	CAMERONMARLOW19770328XKCD	e4b91f9
7	James	Fowler	2/18/1970	XKCD	JAMESFOWLER19700218XKCD	46221bc
⋮	⋮	⋮	⋮	⋮	⋮	⋮
N						

Table: For step 1, each round of the Yahtzee procedure begins with the hashing of the datasets using a new salt. The "Salt" allows us to generate multiple hashes without getting the same hash every round. Next, the hash is divided by the value N/g, where N is the number of individuals in the dataset and $g = 5$ was chosen arbitrarily. The remainder of this calculation is recorded as the group ID. Records are then placed into groups of various sizes based on this group ID. On average the groups should contain g records. Next the frequency of some behavior of interest - in our case voting - is recorded for each group ID. In subsequent steps, a group ID is generated using the identical process on a second dataset. In the second dataset, the frequency of the behavior of interest is assigned to each record based on its group ID. In some cases, the same record is in both datasets, and its contribution to the value assigned to the group in the origin dataset will be transferred to the group in the destination dataset. However, individual records are never matched. We can be sure that identical records in both datasets will be assigned the same group ID, but we can never be sure for any one record if a true match exists in the other dataset or just records that hash to values with the same remainder after dividing by N/g.

The Yahtzee Method: Matching Groups m Times

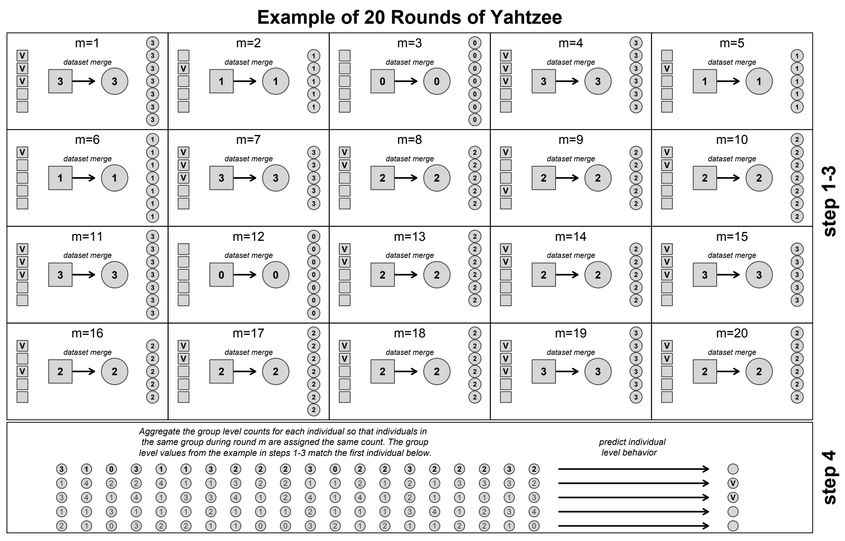

Figure: In step 2 the group ID is determined for all groups where $g = 5$ in the origin dataset and then matched to the same group ID from the destination group-level dataset. Notice that the hashing procedure and group aggregation is the same in both datsets except we keep all groups in the destination dataset, regardless of size. This is so because we only need to know the group size from the origin dataset to make predictions about the behavior in the destination dataset. Once the group-level datasets are matched by the group ID, the group-level information is stored and the process is repeated m times. In step 3 the group level data is sent to the holder of the destination dataset so that the group level values can be assigned to the individual observations based on the same hashes used in the construction of the groups during each of the Yahtzee rounds. Once the destination dataset has acquired a sufficient number of group level values it is possible to then use the combined information to predict the behavior of each individual, which is step 4 of the Yahtzee procedure. For our application, it is possible to predict if the individual is unregistered, a voter or an abstainer. Finally, it is worth repeating that only the group-level data is ever passed from the origin to the destination dataset.

The Yahtzee Method: Selecting m

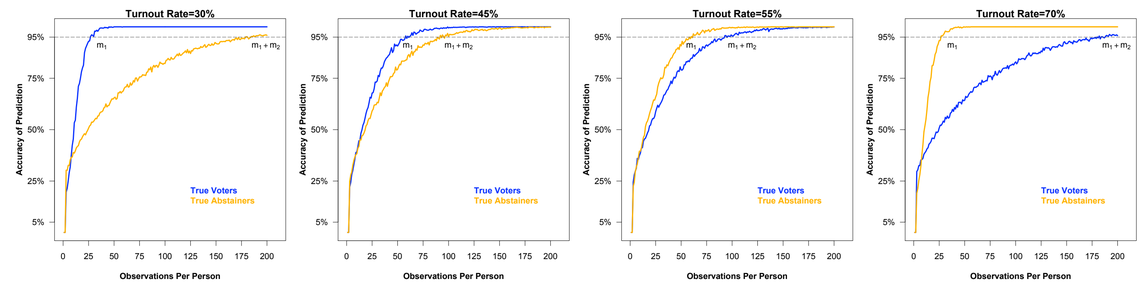

Figure: The proportion of correct predictions for participation rates of 30%, 45%, 55%, and 70% (the match rate is held constant at 30% in all four figures) from a simulation of the matching procedure. The dark line represents the accuracy rate for true participators. The light line represents the accuracy rate for true abstainers. accuracy increases for both categories as observations for each individual are obtained from the Yahtzee procedure. Note that the less frequent of the two behaviors requires fewer observations for classification than the more frequent behavior. m_1 is the number of observations per person necessary to achieve a given level of accuracy for the less frequent behavior and $m_1 + m_2$ is the number of observations necessary to achieve a given level of accuracy for the more frequent behavior.

Application: Matching Voter Registration and Facebook Data

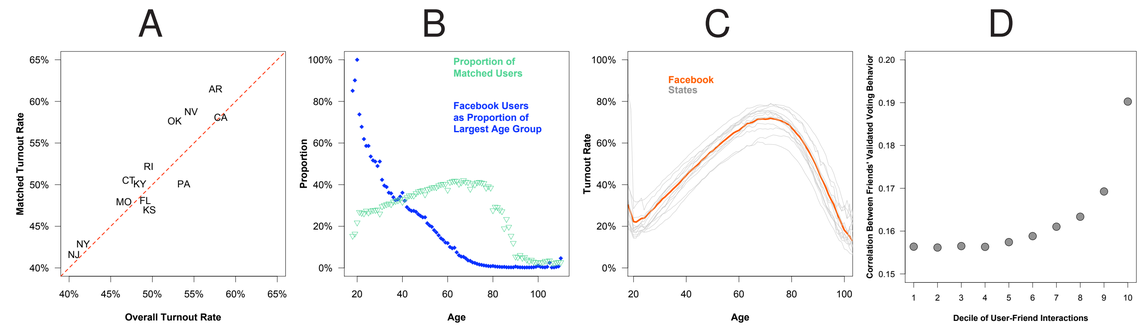

Figure: ***Panel A***: The proportion of matched users who turned out to vote compared to the overall turnout rate by state. The voting rate among Facebook users matched using the Yahtzee procedure correlates highly with the voting rate in the state voting records.
Panel B: The proportion of Facebook users that were matched to the validated voting record by age and each age group's proportion of the largest age group (those 20 years of age at the time of the election). This figure helps to explain why match rates are lower for Facebook users who tend to be younger and more difficult to match than the average registered voter.
Panel C: The proportion of matched users who turned out to vote by age. The dark line represents the turnout rate by age of the matched sample of Facebook users. Each gray line represents the turnout rate by age of a state voter record. The results show that users on Facebook exhibit the same pattern of turnout with respect to age as the populations in other states.
Panel D: The correlation between friends' validated voting behavior based on the proportion of interaction between the dyad in the three months prior to the election. We categorized all friendship in our sample by decile, ranking them from lowest to highest percent of interactions. Each decile is a separate sample of friendship dyads. For example, decile 1 contains all friends at the 0th percentile of interaction to the 10th percentile while decile 2 contains all friends at the 11th percentile of interaction to the 20th, and so on. Interactions include actions on Facebook that could be directed from one user to another and include: comment, like, message, poke, wall post, tag or chat.

Summary of Results

- We validated our method on Facebook and public voter records
- The turnout rate of Facebook users by state strongly correlates with the overall turnout rate of all individuals in the state
- Facebook users within each age group tend to vote at about the same rate as members of those age groups in the population

Conclusion

The results not only suggest that the Yahtzee method works as expected, but also that Facebook users are very similar to the population as a whole in terms of their voting behavior. This is an important finding for researchers who rely on internet websites such as Facebook or Amazon's Mechanical Turk in order to recruit subjects.

Download the Paper

http://arxiv.org/abs/1112.1038