Introduction

We analyze two high-dimensional datasets from the Kinsey Institute for Research in Sex, Gender, and Reproduction. Our goal is to gain an understanding of this type of data and use that knowledge to develop ways to protect it.

Results

Fig. 1 illustrates the differences in survey questions’ and fields’ uniquification rates. Most survey fields can be used to uniquify a high percentage of participants. Fig. 2 shows that, despite high uniquification rates, records are very similar to their nearest neighbor. Participants have very similar nearest neighbors for even unscaled modules. Fig. 3 gives results for experimental record re-identification. Auxiliary information was sampled from within the stated question sets for each record. A larger number of matches indicates a smaller probability of re-identification.

Explanation of Terms

Uniquification: Reduction of population to a single participant
KISBQ (unscaled): Risky sexual behavior
MSQ, STAI, and ZPDR (scaled): Mood and sexuality
SIS/SES (scaled): Sexual inhibition and excitation
Scoreboard: An algorithm to experimentally re-identify participants given auxiliary information.
*: Removing text entry from consideration
†: Participants with more than the stated number of unique combinations or single fields

Acknowledgments & References

This work is supported by NSF grant CNS-1012081.
