Charles Cook; Feb. 28th, 2022

Trustworthy Machine Learning; Project Proposal

“Practical Methods of Inventory Dataset Anonymization & Privacy”

Project Motivation

In virtual marketplaces, particularly that of the Steam network, traders and collectors often keep the visibility of their whole inventory public. This can lead to individuals possessing rare or valuable items being pestered by scammers or hackers, seeking to attain said items through underhanded means. While only a minor nuisance, the only hard option to prevent this negative attention is to set ones inventory visibility to private.

It would be a quality of life improvement if some middle ground existed; That is, if users could have an option of inventory visibility that:

a. Lessens a scammer’s ability to discern if particular items are in the inventory
b. Maintains a trader’s ability to appraise the value of a subset of the inventory

Then a user could still maintain a level of prestige by the worth of their items, or attract good-intentioned trade offers, while deterring scammers and hackers from attempting their schemes.

Project Details

To tighten the scope of how many attributes constitute each item in the inventory dataset, I will curate a collection of 200 to 400 video games for the NES and SNES consoles, and include the Title, Publisher, Release Year, Genre, and Keywords attributes. I will then use techniques of k-Anonymization, wherein certain attributes are obfuscated with a generic value, as a first pass. For a second pass, I will randomly sample the resulting k-Anonymous dataset and calculate the ϵ and δ parameters per the definition of $(\beta, \epsilon, \delta)$ Differential Privacy.

Across different amounts of k-Anonymization (how many attributes are genericized) and different sampling rates (value of β), I will then measure two utility values; One for a scammer, and one for an appraiser:

- The scammer will repeatedly choose a game from the total library of NES/SNES titles, flip a coin, and either calculate if the chosen game is in the set based on the sampled k-Anonymous set, or make a uniform random guess
- The appraiser will compute summary statistics from the sampled k-Anonymous dataset, and combine these with individual valuations of games to extrapolate an estimated total worth of the original dataset

I will seek to determine at which combination of k-Anonymization and $(\beta, \epsilon, \delta)$ Differential Privacy strike a balance between lessening the scammers accuracy while heightening the appraiser’s valuation.

Existing Work & Bibliography

[1] Ninghui Li, Wahbeh Qardaji, and Dong Su. 2012. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security (ASIACCS ’12).* Association for Computing Machinery, New York, NY, USA, 32–33. DOI:https://doi.org/10.1145/2414456.2414474

[2] Latanya Sweeney. 2002. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5 (October 2002), 571–588. DOI:https://doi.org/10.1142/S021848850200165X

[3] Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II (ICALP’06).* Springer-Verlag, Berlin, Heidelberg, 1–12. DOI:https://doi.org/10.1007/11787006_1