Controlling privacy risk in database studies for human subject protection(HSP) via a privacy budgeting system

Hye-Chung Kum[1,2,3], Cason Schmit[1], Alva O. Ferdinand[1], Theodoros Giannouchos[1]

[1] Texas A&M University Health Science Center, Department of Health Policy & Management, [2] Texas A&M University, Department of Computer Science & Engineering
[3] Population Informatics Research Group (https://research.tamhsc.edu/pinformatics/)

Background

- Recent advances in information privacy have shown the need to think about ethical data analysis as a budget-constrained problem.
- The goal is to quantify and contain privacy risk under a fixed privacy budget appropriate for the purpose.
- A well-balanced system would allow maximized benefit while maintaining research risks under a fixed budget.

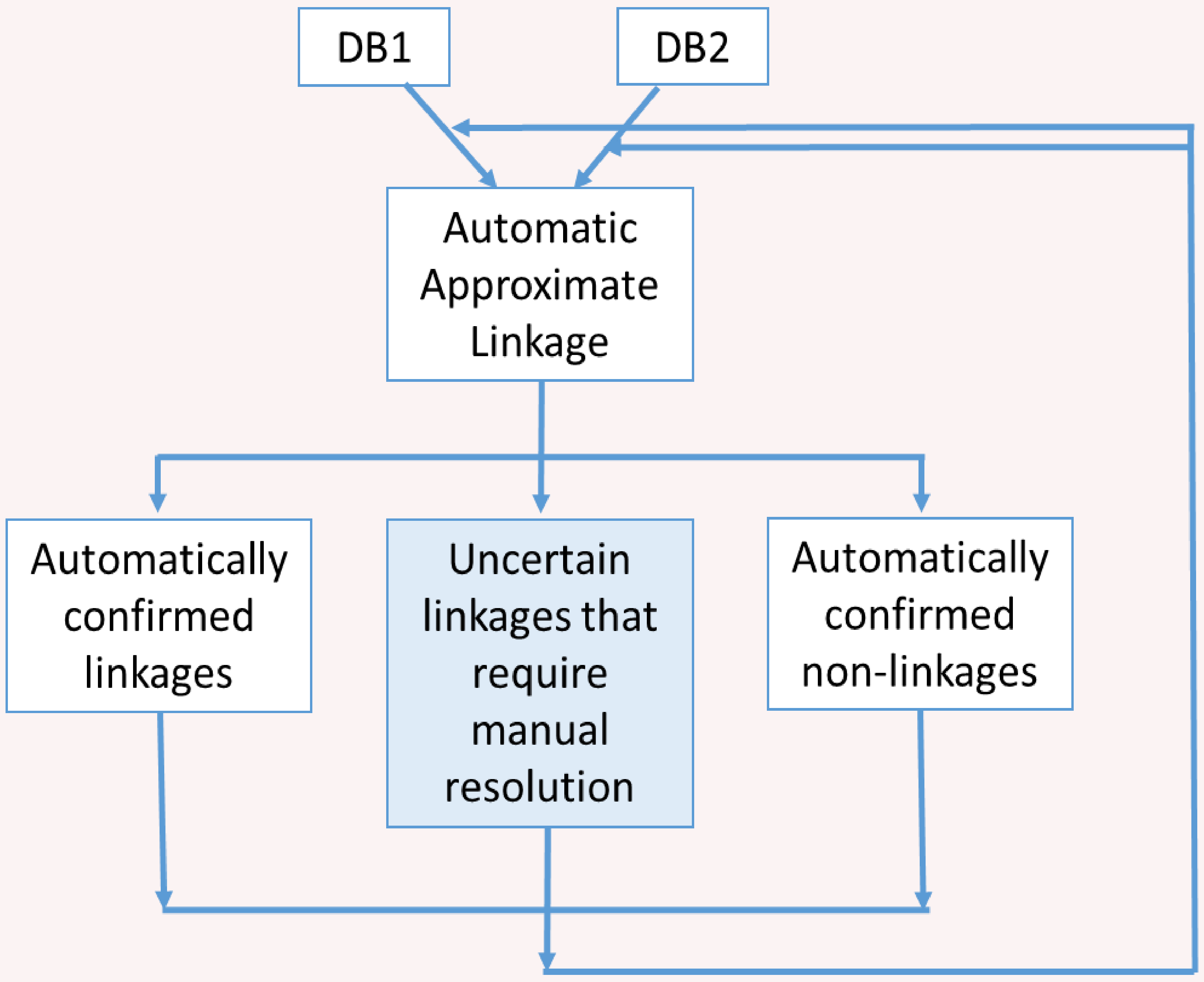
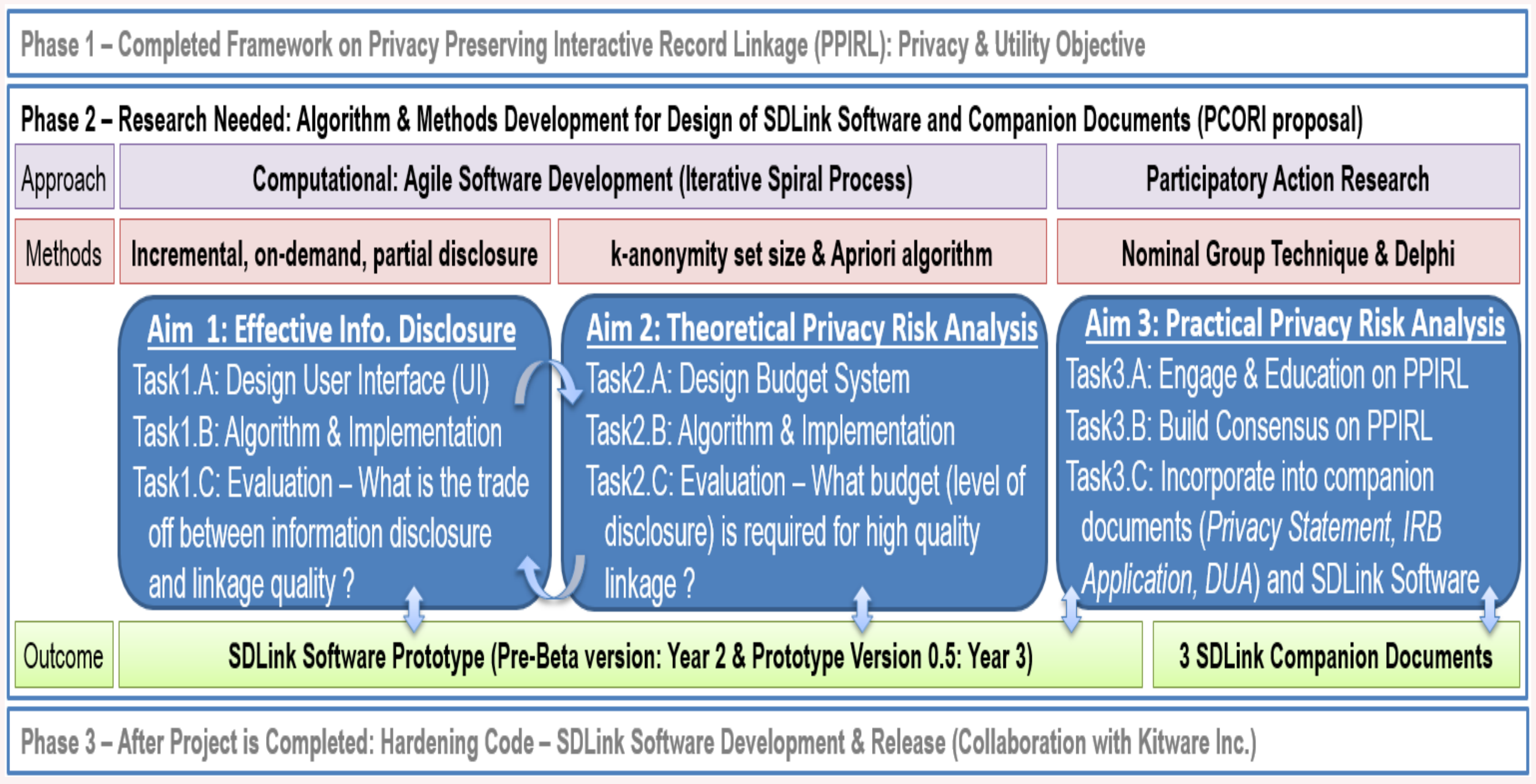

DB1
DB2
Automatic Approximate Linkage
Automatically confirmed linkages
Uncertain linkages that require manual resolution
Automatically confirmed non-linkages

Figure 1 - Approximate RL Human-computer system

Methods

- We propose a privacy budget system for human subject protection that uses anonymity-set size to define allowable risk in database studies.
- The anonymity-set size, n, is the number of people in the population that share the same identifying information disclosed during research which can quantify the risk of identity disclosure.


Phase 1 – Completed Framework on Privacy Preserving Interactive Record Linkage (PPIRL): Privacy & Utility Objective

Phase 2 – Research Needed: Algorithm & Methods Development for Design of SDLink Software and Companion Documents (PCORI proposal)

Approach: Computational: Agile Software Development (Iterative Spiral Process) | Participatory Action Research

Methods: Incremental, on-demand, partial disclosure | k-anonymity set size & Apriori algorithm | Nominal Group Technique & Delphi

Aim 1: Effective Info. Disclosure
Task1.A: Design User Interface (UI)
Task1.B: Algorithm & Implementation
Task1.C: Evaluation – What is the trade off between information disclosure and linkage quality ?

Aim 2: Theoretical Privacy Risk Analysis
Task2.A: Design Budget System
Task2.B: Algorithm & Implementation
Task2.C: Evaluation – What budget (level of disclosure) is required for high quality linkage ?

Aim 3: Practical Privacy Risk Analysis
Task3.A: Engage & Education on PPIRL
Task3.B: Build Consensus on PPIRL
Task3.C: Incorporate into companion documents (*Privacy Statement, IRB Application, DUA*) and SDLink Software

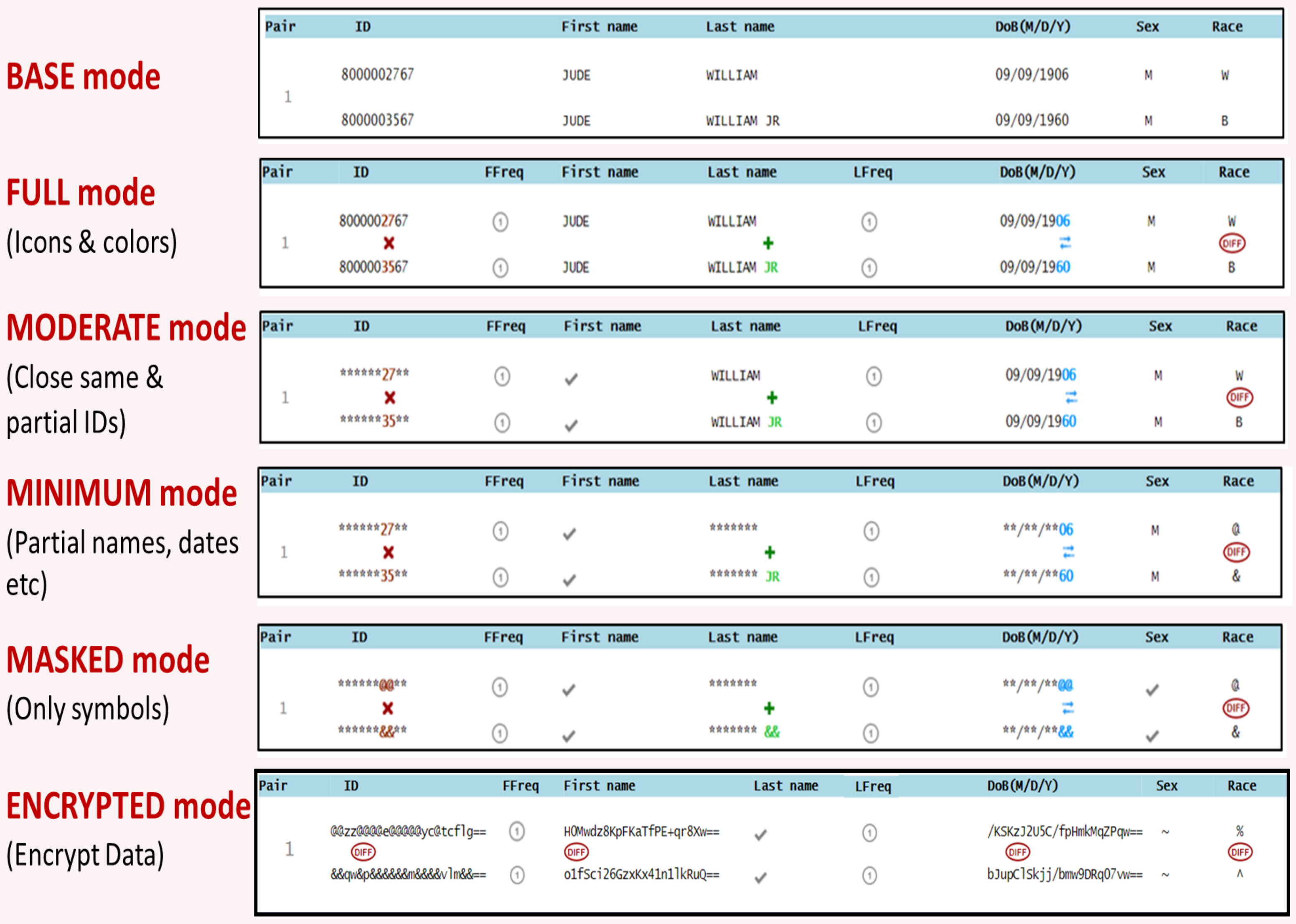
Outcome: SDLink Software Prototype (Pre-Beta version: Year 2 & Prototype Version 0.5: Year 3) | 3 SDLink Companion Documents

Phase 3 – After Project is Completed: Hardening Code – SDLink Software Development & Release (Collaboration with Kitware Inc.)

Figure 2 – PPIRL Software Development Phases

Discussion

- Traditional mechanisms for HSP (e.g., meaningful informed consent, de-identification) improperly prioritize the ethical principle of autonomy by restricting research that advances the principles of beneficence and justice.
- In recognition, recent changes to the Common Rule better facilitate database studies by simplifying secondary data analysis HSP requirements. This privacy budget system aligns with the new Common Rule and promotes a better understanding and quantifying of specific privacy risks in research.
- Novel methods for HSP based on digital technology that can facilitate better reasoning, communication, and negotiations of privacy risk in database studies can improve transparency in database studies ultimately improving human subject protection.


BASE mode

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
1	8000002767	JUDE	WILLIAM	09/09/1906	M	W
	8000003567	JUDE	WILLIAM JR	09/09/1960	M	B

FULL mode
(Icons & colors)

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	8000002767	①	JUDE	WILLIAM	①	09/09/1906	M	W
	✗			+		⇄		DIFF
	8000003567	①	JUDE	WILLIAM JR	①	09/09/1960	M	B

MODERATE mode
(Close same & partial IDs)

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	******27**	①	✓	WILLIAM	①	09/09/1906	M	W
	✗			+		⇄		DIFF
	******35**	①	✓	WILLIAM JR	①	09/09/1960	M	B

MINIMUM mode
(Partial names, dates etc)

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	******27**	①	✓	*******	①	**/**/**06	M	@
	✗			+		⇄		DIFF
	******35**	①	✓	******* JR	①	**/**/**60	M	&

MASKED mode
(Only symbols)

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	******@@**	①	✓	*******	①	**/**/**@@	✓	@
	✗			+		⇄		DIFF
	******&&**	①	✓	******* &&	①	**/**/**&&	✓	&

ENCRYPTED mode
(Encrypt Data)

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	@@zz@@@@e@@@@@yc@tcflg==	①	HOMwdz8KpFKaTfPE+qr8Xw==	✓	①	/KSKzJ2U5C/fpHmkMqZPqw==	~	%
	DIFF		DIFF			DIFF		DIFF
	&&qw&p&&&&&&m&&&&vlm&&==	①	o1fSci26GzxKx41n1lkRuQ==	✓	①	bJupClSkjj/bmw9DRq07vw==	~	^

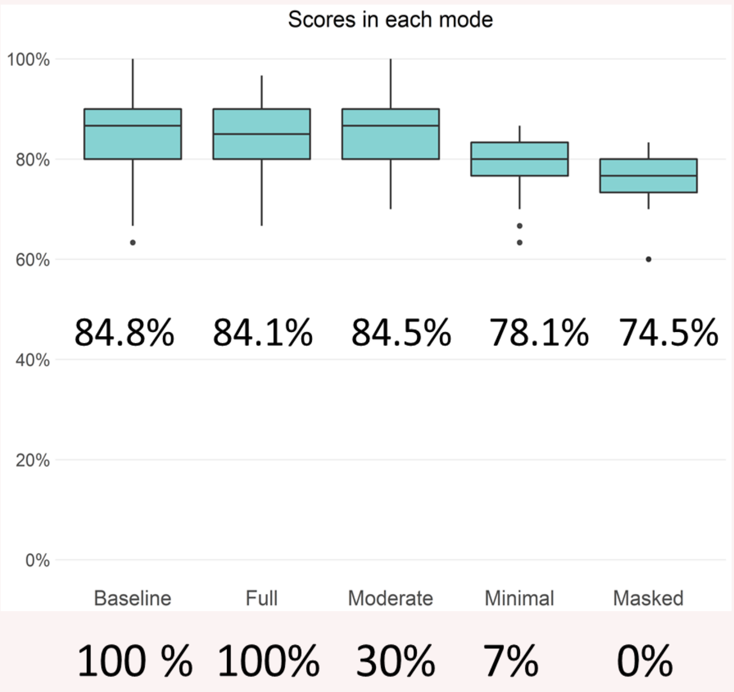
Figure 3 – Disclosure Modes

Controlling privacy risk in database studies for human subject protection(HSP) via a privacy budgeting system

Hye-Chung Kum[1,2,3], Cason Schmit[1], Alva O. Ferdinand[1], Theodoros Giannouchos[1]

[1] Texas A&M University Health Science Center, Department of Health Policy & Management, [2] Texas A&M University, Department of Computer Science & Engineering
[3] Population Informatics Research Group (https://research.tamhsc.edu/pinformatics/)

Results



- We can get comparable results to full mode with only 30% disclosure with appropriate masks (moderate mode)
- As we mask more values for privacy, quality of results start to suffer (p<0.001)
- However, even legally de-identified data with proper masks can be linked properly for most situations
  - 0% disclosure still had 75% accuracy
  - Incremental disclosure can significantly improve privacy protection with negligible impact on quality of linkage

Figure 4 – Accuracy Score by Disclosure Mode

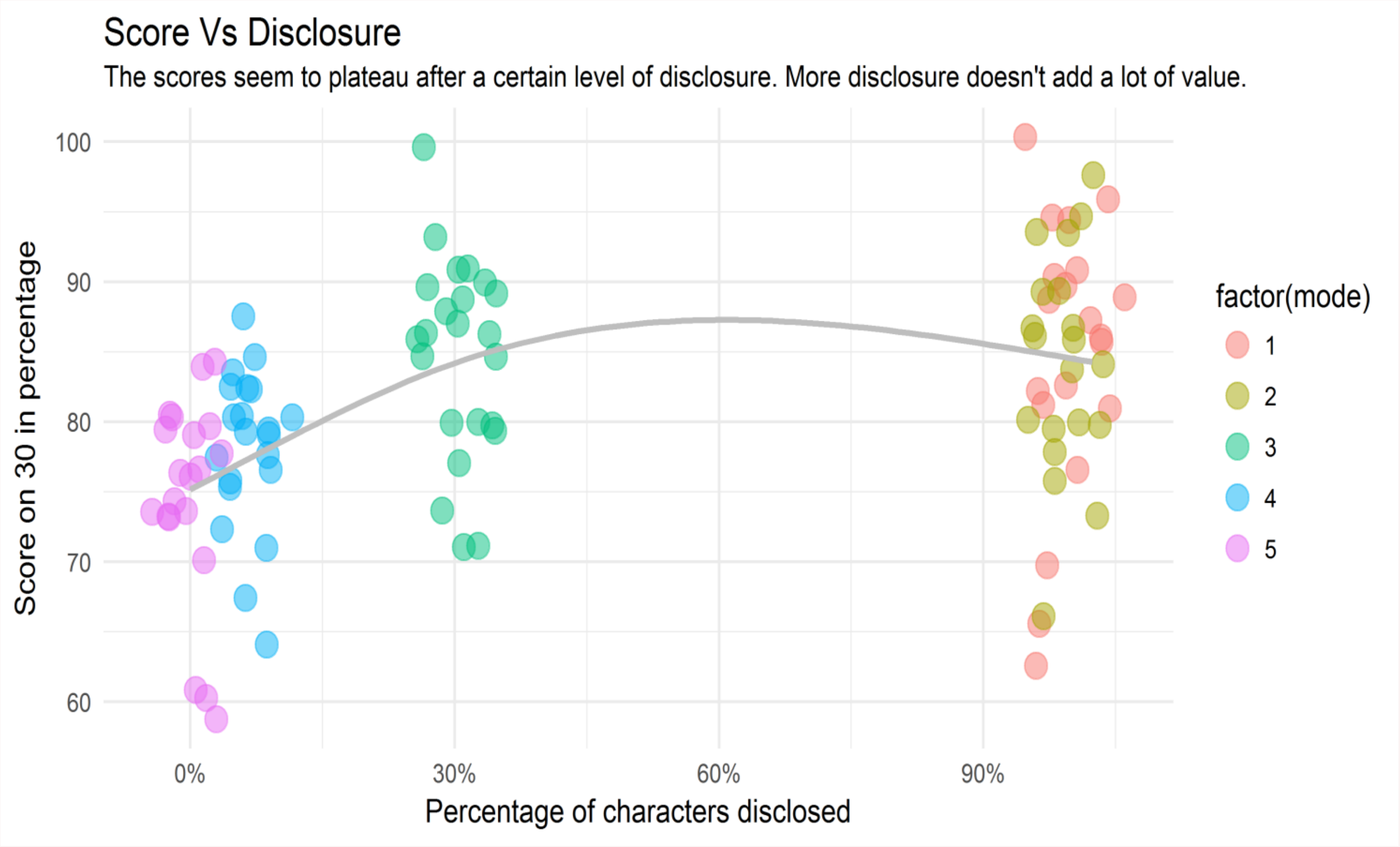


Figure 5 –Score vs Distance

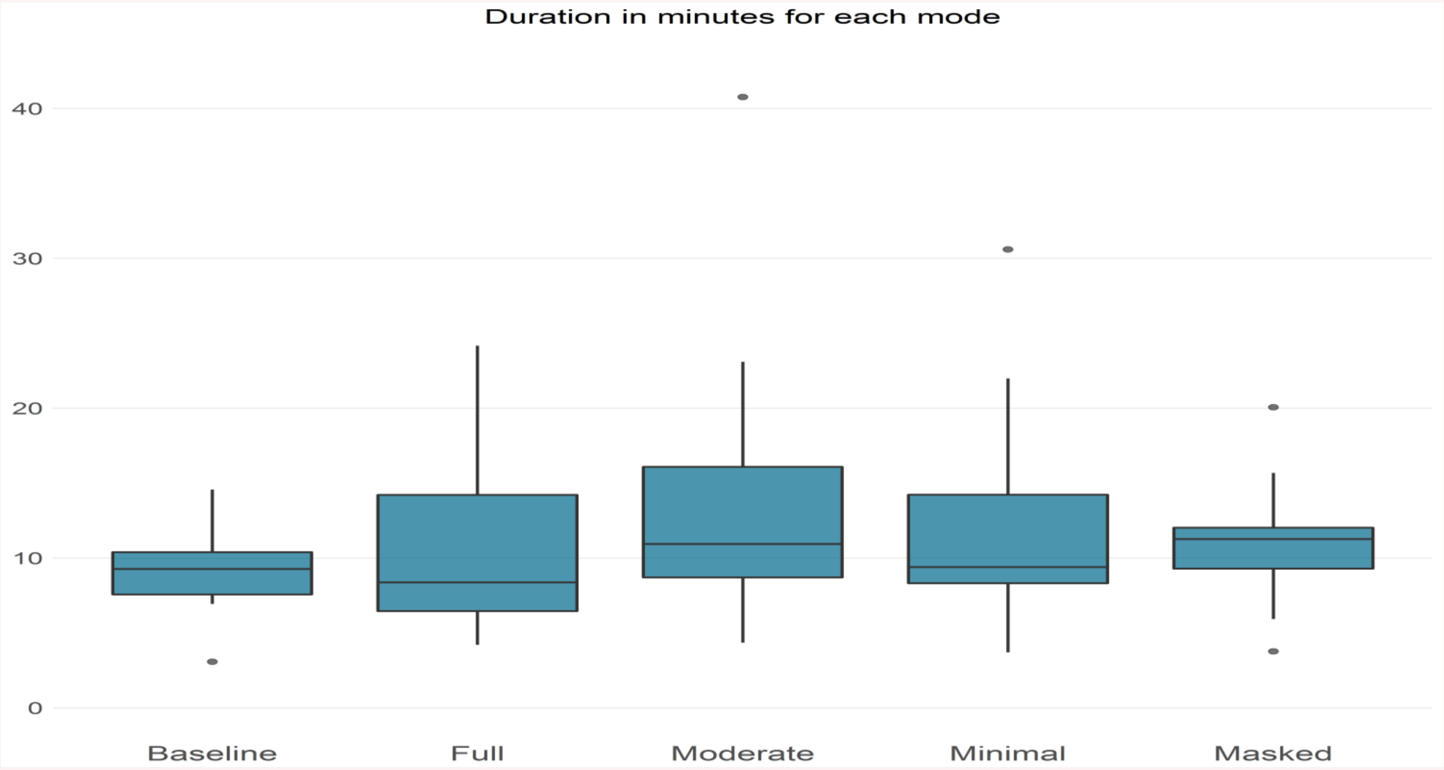


Figure 6 –Time by Disclosure Method

Conclusion

- Mathematically, the identity disclosure risk is inversely related to the number of entities in the population that share the information disclosed during research.
- If the information refers to one and only one person in the population, then the identity of the person has been fully disclosed by the information revealed.
- On the other hand, if the information disclosed is identical for multiple people (=n), referred to as the anonymity-set size, then the information is less revealing as it could refer to any one of the n people.
- For example, when a frequently occurring name (e.g., "Mary") is disclosed, there is low probability that the identity of a specific person named Mary is revealed. In comparison, when a rare name (e.g., "Hye-Chung") is disclosed, there is a greater risk that the identity will be ascertained.
- By combining all anonymity-set sizes for different information disclosed during a given study, we design a privacy risk score based on the background population.
- When the background population information is not readily available, as is often the case, the study population can be used to calculate the most conservative privacy risk score.
- The proposed privacy budget system can enable all stakeholders (e.g., database researchers, IRBs, the public) to (1) accurately reason about (2) communicate on, and (3) agree on the acceptable quantified privacy risk considering the anticipated benefits of a given study.
- Ideally, the budget would be a simple measurement on the probability of identity disclosure for the population. It will also allow for better monitoring of risk to increase transparency.

Limitations

- The proposed method has not been implemented and no evaluation is currently planned.
- More time is required to finalize and evaluate the concept.
- This poster provides a great opportunity to gather wide input from the IRB community on the novel concept before the privacy budget design is finalized and implemented.