Why a Systems-Science Perspective is Needed to Better Inform Data Privacy De-identification Public Policy, Regulation and Law

Daniel C. Barth-Jones, Ph.D.
Assistant Professor of Clinical Epidemiology
Mailman School of Public Health
Columbia University

Systems sciences pursue understanding of how parts (including individual actors and groups) within a larger system interact to produce emergent phenomena which are greater than, or different from, a mere sum of the parts. Important examples of systems behaviors include the existence of threshold phenomena, which allows infectious diseases to spread as epidemics through a population, or the sudden crystallization of supersaturated solutions. I argue that data privacy policy for de-identification must take a systems perspective in order to better understand how combined multi-dimensional technical and regulatory interventions can effectively combine to create practical controls for countering wide-spread re-identification threats.

In his seminal "*Broken Promises*" work on anonymization, Paul Ohm puts forth a dystopic vision contending that the failure of perfect de-identification will lead all of us through inescapable cycles of accretive re-identifications toward our personal "databases of ruin". Narayanan and Shmatikov extend Ohm's argument by further contending that "any attribute can be identifying in combination with others". Such conceptions ignore the fact that data uniqueness, replicability and accessibility all contribute importantly to the probability of realizing successful data re-identifications. By fallaciously conflating "what is possible" with "what is probable", Ohm's assertion of strictly accretive re-identification ignores important underlying mathematical realities regarding information entropy and signal detection theory. However, when statistical disclosure limitation methods have been used to de-identify data, it is frequently possible to practically achieve orders of magnitude reductions in data re-identification probabilities, resulting in dramatic increases in "false-positive" re-identifications. Because false-positive data linkages will typically add incorrect data into electronic dossiers, they can help prevent strictly accretive re-identification, thus disrupting Ohm's supposed systemic "crystallization" of iteratively linked de-identified data into accurate dossiers for the vast majority of a population. While human rights based views of data privacy lead us to value the data privacy of all individuals, de-identification's lack of perfection cannot justify its wholesale abandonment if, on a systems level, it can be shown to be capable of preventing mass re-identification as part of multi-dimensional regulatory approaches. Systems modeling and quantitative policy analyses, including uncertainty analyses, provide us with the necessary scientific tools to critically evaluate the potential impacts of pseudo/anonymization in various regulatory schemas and should be pursued vigorously as a routine approach for conducting data privacy policy evaluations.