Title: ***Data privacy in an age of increasingly specific and publicly available data: An analysis of risk resulting from data treated using Anonos' Just-In-Time-Identity Dynamic Data Obscurity methodology.***

Author: Sean Clouston, PhD, #3-096, Stony Brook University, Health Sciences Center, 101 Nicholls Rd., Stony Brook, NY, 11794. Tel. (631) 444-6593, fax: (631) 444-3480, e-mail: sean.clouston@stonybrookmedicine.edu.

Abstract: The public's interest in the reasonable expectation of privacy is met when personal data (PD) remains private: within reasonable limits, the data cannot be used to single out, or to inferentially identify or link PD to a particular data subject. Historically, privacy has been maintained by reducing access to identifiable data, while ensuring that the likelihood of re-identification, largely interpreted through equivalence classes, is reduced. However, PD is increasingly measured outside of traditional situations, contains increasingly detailed information, and is being utilized by new entities in new purposes. Yet, the public maintains the same interest in privacy. As a result, static de-identification tactics, which delete known identifiers by replacing them with one pseudonym used consistently for an individual, have been increasingly susceptible to critique. One way forward is to use new temporally dynamic obscurity protocols that actively minimize the risk of re-identification. This report analyzes the ability of dynamic data obscurity to minimize re-identification under different circumstances and thus to ensure data privacy. Analyses provided in this report aid in assessing privacy and security risks and maintaining privacy and security when data incorporates detailed and even longitudinal information. PD is kept private within an acceptable level of risk, subject to some constraints on oversight and sample size. Data stored or transmitted pursuant to these methods is de-identified in the traditional sense; further, data storage and transmission are more robust using these methods for a number of reasons outlined in the report. Moreover, because data security is dynamic, privacy policies can be flexibly implemented to ensure security is consistently and completely ensured.

Conflicts of interest: This report was originally written with Dr. Clouston acting as an external and independent consultant for Anonos, at which time the author had no conflicts of interest to report. However, during the process of the finalization of this report, but before publication, the author became an Anonos shareholder. Anonos played no role in the study conclusions, and did not direct the analysis. Anonos provided editorial suggestions relating to the implications and implementations of the analytic results provided herein.

Introduction

Data privacy is important for many reasons. As just one example, health data, once released, could be used to reveal realities about a person's health that he or she may not want to share, including diagnoses of societally stigmatized diseases (e.g., PTSD, HIV/AIDS, Schizophrenia, etc.) and health issues having financial implications for their families or their health insurance carriers (e.g., physical activity, blood pressure, financially discriminatory actions, etc.). On the other hand, as Matthews and Harel (2011) highlight in their review of data privacy issues and solutions, researchers must ensure that data are accessible for potentially valuable research applications. Moreover, data are increasingly collected as byproducts of private sector innovation, but while these data need to be protected, it is not in the public interest to stifle innovation requiring that data. Static de-identification, which we here define as de-identification that maintains the structure and linkages of data and achieves de-identification by replacing identification data with a randomized static pseudonym, was previously deemed to be sufficient for reducing the risk of re-identification. However, new data, more powerful types of data analytics, and the increasing number of data sources have made researchers, policymakers, and software developers skeptical this can continue (Chen & Zhao, 2012; de Montjoye, Radaelli, Singh, & Pentland, 2015).

As one example of how regulations are affected by the issues surrounding data minimization, a recent U.S. Federal Trade Commission report notes that while HIPAA traditionally only pertains to a small number of people handling health information, such as doctors or hospitals, "health apps are [increasingly] collecting this same information through consumer-facing products, to which HIPAA protections do not apply…" and goes on to state that "consumers should have transparency and choices over their sensitive health information,

regardless of who collects it" (Federal Trade Commission, 2015). The conclusion of the FTC report was twofold: the majority decision supports the need for "data minimization," or the wholesale deletion of collected information from the information ecosystem, while the minority decision notes that this data minimization might negatively impact health-related research and decision making (Federal Trade Commission, 2015).

The FTC dissent highlights the contrast between data value and data privacy. A non-partisan research firm (the Information Technology and Innovation Foundation or ITIF), further highlights problems with data minimization in the private sector: "the FTC's report correctly recognizes that the Internet of Things offers potentially revolutionary benefits for consumers and that the industry is still at an early stage, [but the report] unfortunately attempts to shoehorn old ideas on new technology by calling for broad-based privacy legislation"; further, "in calling for companies to reduce their use of data, the FTC misses the point that data is the driving force behind innovation in today's information economy" (Castro, 2015). These dissenters each view such data collection and analysis efforts as *serving the individual and public interests*, even at the cost of privacy. *Wired* magazine concretizes these dissents, reporting that though IoT devices currently being developed are geared towards gathering "reams of largely superficial information for young people whose health isn't in question, or at risk" (Herz, 2014), "the people who could most benefit from this technology—the old, the chronically ill, the poor—are being ignored… [primarily because] companies seem more interested in helping the affluent and tech-savvy sculpt their abs and run 5Ks than navigating the labyrinthine world of … HIPAA."

Three Main Limitations

There are three main limitations with the current approach data privacy. First, static de-identification is not robust. Second, transmission is particularly problematic. Third, an increasing

number of entities are involved in providing guidance about privacy in a way that is increasingly difficult to maintain. These are discussed in detail in the following section.

First, data are not truly de-identifiable. Specifically, a recent article in *Science* observed the relative ease with which one can uniquely identify individuals using only small amounts of financial information (de Montjoye et al., 2015): indeed, for 90% of the cited sample only four pieces of information were needed to achieve "unicity" – i.e., development of unique identifying profiles derived from traditionally de-identified financial data. As noted by (El Emam, 2015), unicity in the dataset does not mean that any person has successfully re-identified each individual; however, once de-identified and made available to the public, data are subject to "data fusion", which is the linking of multiple different datasets together in order to broaden our understanding of the people in a dataset. The risk of data fusion has lead to this finding being highly publicized, for example the *Wall Street Journal* noted that unicity in financial data meant one could readily "find the name of the person in question by matching their activity against other publicly available information such as LinkedIn and Facebook, Twitter, and social-media check-in apps such as Foursquare" (Hotz, 2015). The *Harvard Business Review* concludes the implications of this work "are profound. Broadly, it means that static anonymity doesn't ensure privacy" (Berinato, 2015).

Second, current data de-identification tactics when data are being transmitted are especially questionable, such transmission increasingly occurring through devices considered to be within the "Internet of Things" (IoT) (Rivera & van der Meulen, 2014). During transmission, the normal "size" of the dataset is curtailed, further weakening the assumptions on which we rely for data security. At the same time, the amount and specificity of data is increasing with data available such as the person's everyday progression from home to work or the number of

calories, types of food, and restaurants that they ate in during their last week. Furthermore, IoT devices increase the importance of information transmission; for example, in the case of healthcare information, clinicians might be able to use interconnected devices to monitor a patient's health, including vital signs or physical activity, potentially raising new concerns regarding data privacy and an increased risk of data breach (Tyrrell, 2014).

Finally, de-identification, proceeding in a static manner, must be implemented under one specific policy regime to the detriment of others. For example, it may be that data collected under one policy are made more or less secure than are necessary under a newer or different structure so that data managers either must redo their work to a different standard, resulting in substantial inefficiency, or may simply choose not to allow access to data because the cost is too large to ensure compliance. In such a circumstance, having pre-approved levels of access could help to ensure that data are both accessible to individuals from varying regions or policy regimes, and at varying levels of security.

Data Construction

Without a solution that responds to these concerns and truly de-identifies PD, a broad range of individuals including, but not limited to, software developers and information technology specialists, will have access to non-de-identified PD data. Static de-identification, as noted above, is not working. Dynamically obscuring data may be one way to retain data privacy while reducing the risk involved in collecting, storing, and analyzing such data (Warren, 2014). Examining new ways requires a basic knowledge of dataset construction, different types of data, and existing de-identification protocols. The following sections provide an introduction to topics underlying these issues before moving on to a more formal analysis of the risk of re-identification.

De-identification

The process of de-identification decreases privacy risks to individuals by removing identifying information from protected or personal data (PD). Thus, in the above dataset, data from the first column (identifying information, here an IP address) would need to be removed from the dataset. We should note, however, that de-identification usually references two somewhat separate processes: the removal of the certainty that any particular individual is part of the observed dataset, and the removal of the certainty that any particular observation might, in the correct circumstances, be sufficiently unique to be re-identified with other available data. Thus, while it is often believed that removing these IP addresses renders similar datasets (usually with more observations) "de-identified" in the traditional sense, as discussed above many of these observations can be uniquely identified using data characteristics that can lead to "unicity" within the database, rendering them unique in the data and thereby at risk of re-identification (de Montjoye et al., 2015).

The problem of de-identification has historically been addressed in a number of temporally static ways. Matthews and Harel (2011) list the following techniques used for de-identification: 1) limitation of detail, 2) top/bottom coding, 3) suppression, 4) rounding, 5) adding noise, and 6) sampling. *Limitation of detail* works through categorizing or collapsing information to reduce the possibility of characteristic re-identification. *Top/bottom coding* characterizes the replacement of observational data with a "top-code", an upper limit on all published values of a variable, and/or a "bottom-code", a lower limit on all published values of a variable, the replacement of which reduces the likelihood of re-identification of data that are more likely to be unique, such as very high incomes, by recoding them so that outlying observations are grouped together. *Suppression* works by removing potentially identifiable data

from the publically available dataset. *Rounding* introduces noise by randomly re-assigning rounded values to all the individuals in a dataset, and is often used for ages because though we may know that individuals are 45.67 years old (i.e., 45 years and 8 months), we recode that information into yearly (as age 45) or into age groupings (such as 45-49). *Addition of noise* uses a randomization routine to change the values in each cell by some random amount, an approach often used with geographic data such as that in the *Demographic and Health Surveys*, which have randomly dispersed geographic residential locations by some distance less than five kilometers (Measure DHS & ICF International, 2013). Finally, *sampling* resolves de-identification by requiring that data released be only a subset of the data available, with the convention that between 95-97% of the data collected could be released; however, sampling also resolves another issue, notably that individuals should not be known to have been a part of the dataset, because it removes, at random, entire individuals from a dataset so that you may not be certain that any particular person who was originally contained within the dataset are also contained within the dataset released.

Since then, more complex mathematical routines have been used to ensure data is kept confidential and that this confidentiality is unlikely to be broken. The most useful of these build on the randomization approach because it is the most secure and removes the least value from the data. *Matrix masking*, for example, codifies the data by multiplying them by a form of encryption key that researchers must know about and account for when analyzing data (Cox, 1994). Another particularly interesting method, called *synthetic data,* uses data matching methods originally built to provide results for missing data to effectively swap characteristics between individuals in a dataset, thereby retaining some of the statistical uniqueness and value while reducing the risk

of re-identification (Rubin, 1993). Note that these methods work by *increasing uncertainty* about the identity of any one individual without unnecessarily modifying the data itself.

De-identification

Current static de-identification methods generally rely on data being purposefully collected into one dataset before being anonymized, using usually one or two of the previously mentioned techniques, to protect privacy. These datasets are then shared with the public either at large or through limited access policies depending, largely, on the level of specificity provided in the data, or kept hidden from the public entirely.

This data fusion may be increasingly easy to do, especially as individuals "check in" on publicly available social networking sites. These data fusion techniques are big, flexible, fast, constantly changing, and being made more specific; they also use data that are being increasingly used, held, or transmitted by an ever larger number of individuals. Thus, the conclusions of de Montjoye et al. (2013, 2015) are likely true: if data are unique, those data may be readily identifiable among a substantial portion of users.

Estimating risk of re-identification

The probability of re-identification can be estimated (El Emam, Dankar, Vaillancourt, Roffey, & Lysyk, 2009). Re-identification requires access to the full datasets, which contain information for six separate individuals, and that re-identification is being undertaken purposefully to find a particular individual who may be in the dataset. We provide an example of a blood-pressure monitoring application on a internet-connected wearable device (i.e., a phone, watch, shoe, or other wearable device), which 1) identifies a user, 2) monitors health information, and 3) is linked to geographic positioning systems (GPS) data. Because this is a useful and clearly risky subject, we will rely on this example throughout this analysis. To protect

the privacy of individuals associated with these data, we are faced with the following difficulties: de-identification may be subject to difficulties in both data storage and in data transmission, and further may propose a risk to multiple governing bodies since it collects data that may include health information, could easily integrate different types of data including longitudinal data, and may also have clinical applications if clinicians are interested in monitoring patients' everyday health in this way.

Below we define a data matrix or dataset ***H*** (Table 1), which for simplicity is a 6 x 6 matrix that contains 36 unique data points (called cells). Different rows may contain different information for the same individual if that person is followed over time or is observed by different people (in longitudinal or higher-dimensional data). Note that there are therefore both explicit and implicit identifiers within most datasets: the IP address is explicit while the row number, when not longitudinal data, is an implicit identifier.

Table 1. Hypothetical dataset (***H***) collected from multiple smartphones on the same network by a blood pressure application in December 2014

IP address	Latitude	Longitude	AgeSex	High blood pressure
192.168.0.0	40.13	-79.85	32 M	No
192.168.101.201	40.13	-79.86	45 F	Yes
192.168.4.57	40.15	-79.54	39 M	No
192.168.75.122	40.10	-79.59	45 F	No
192.168.22.40	40.29	-79.54	56 M	No
192.168.1.220	40.29	-79.56	22F	No

Note: IP addresses are assumed static for period of observation; Latitude and Longitude are "rounded off" to two decimal places and so are not precise to the level of a specific house.

For ease, we may label this person *Alice* and, using the example in Table 1 above, suppose *Alice* is the person with high blood pressure corresponding to IP address 168.192.101.201. Let us assume we know the type of information a neighbor, coworker, or employer might know about *Alice*. Suppose, for example, we know she is female, that she was born approximately 40-50 years before the data were collected, that we know that she lives in Belle Vernon, Pennsylvania (Latitude, Longitude = +40.13, -79.85). However, we want to know further whether *Alice* has high blood pressure, and thus we also need to re-identify her using the data provided.

We follow previous reviews of re-identification risk assessment (El Emam et al., 2009) that define an "acceptable risk" as one that is at most $\rho = 0.20$; further, for reasons that will become clear later, we further clarify that an "unacceptable risk" is one known to be greater than $\rho = 0.20$. Then, the whole dataset's risk of re-identification (r) can be defined as: $r = 1/\min_j(f_j)$, where f is the number of individuals with equivalent characteristics to any one particular person, including here *Alice*, and min_j is the minimum number of individuals in a subset of categories (j; sometimes called an equivalence class, the basis for the "unicity" argument) that fit *Alice*'s known characteristics.

The risk has also been specified in the following way, including the measurement of how many equivalence classes that there are in a dataset who are expected to be distinct. Along those lines, Benitez and Malin (2010) provide the following definition of total risk: $r_g = \frac{k}{n} b^{1-n}(b^n - (b-1)^n)$, where k references the number of individuals in b possible equivalence classes for a sample of size n. Because the total risk revolves around the risk within a particular equivalence class, we thus begin by briefly overviewing equivalence classes in re-identification, before explaining why data parsing helps secure privacy.

Re-identification in practice

In table 1 above, our first efforts would be to rely on *Anonos* to mask and eliminate IP addresses, replacing that information with *Anonos*-provided DDIDs. Prior to this, the risk of identification is perfect: $r = 1/1 = 1$. This is an unacceptable risk because $r = 1 > \rho = 0.20$. However, after removing IP addresses, the risk is reduced because we cannot rely on identifiable information. In particular, the risk becomes $r = 1/n = 1/6 = 0.17 < 0.20$, an acceptable risk. We still, however, know that *Alice* is a woman aged 40-50 who lives in Belle Vernon, Pennsylvania. The risk of re-identification as a woman is defined as the inverse of the number of people in the equivalence class: in this case, the inverse of the number of women in the dataset, and thus $r = 1/number_{femalesf} = 1/3 = 0.33$. Because 0.33 is larger than 0.20 (as defined above by our categorization of acceptability, ρ), we note that this is already an unacceptable level of risk. However, for clarification as to the nature of data linkages we push this data further to use more characteristic and specific data. We examine the data and note that there are two women aged 40-50 in the data. Therefore we calculate r = ½ = 0.50; since r > $\rho = 0.20$ we define this as both higher, and an unacceptable risk. We further know that *Alice* lives in Belle Vernon (latitude ranges from 40.12 to 40.14, longitude ranging from -79.84 to -79.86). This shows us that there are two people in these data living in Belle Vernon, and thus we calculate r = ½ = 0.50; since r > $\rho = 0.20$ we define this as an unacceptable risk. Linked together, we can further see that, of those people living in Belle Vernon, only one is a female aged 40-50. Thus, data linking increases our risk of re-identification to $r = 1$, an unacceptable risk resulting in certain re-identification (i.e., "unicity" (de Montjoye et al., 2015)).

Table 2. Hypothetical DDO data (***H***) collected from multiple smartphones on the same network by a blood pressure application in December 2014

DDID	IP address	DDID	Lat.	DDID	Long.	DDID	Age	DDID	Sex	DDID	High BP
5657	192.168.4.57	4858	40.10	5049	-79.86	4958	22	5141	F	6878	No
5854	192.168.101.201	5934	40.13	3060	-79.85	3938	32	5161	F	4948	No
3938	192.168.0.0	3030	40.29	9090	-79.54	4040	45	6236	M	3031	No
5910	192.168.1.220	1234	40.13	8974	-79.54	5010	45	7747	M	4094	No
2039	192.168.22.40	1410	40.15	9030	-79.56	7079	56	8585	M	0967	No
5050	192.168.75.122	4040	40.29	9283	-79.59	7078	39	9999	F	0847	Yes

Note: We have shortened title names to limit table size. Lat.: Latitude; Long.: Longitude; BP: Blood Pressure; DDID: Dynamic de-identifiers. Each DDID references the value in the column to its right. For ease of reference, the above table is represented to include both DDIDs and corresponding data; in an actual Anonos implementation, this table would not contain JITI keys that would be stored in a separate table administered by the Anonos engine directly.

Removing information from the implicit linkages is novel because it removes the possibility of linking data together to contextualize information. Thus, data are not saved in Table 1 above, but in a way that more closely resembles Table 2 above. For ease, the following discussions reference the "worst case scenario," wherein an entire dataset with small sample size is observed. We have here kept the variables in the same order: as would be obvious to any knowledgeable party, each column references similar data within variables and different data between variables; and further, the order of variables is rarely meaningful. Also, four-digit numerical characters were used as DDIDs for simplicity alone; this referencing does not reflect the method through which Anonos derives or defines its DDIDs. We assume, for conservative estimates, that we know which indicator each observation references.

Using this method, both explicit and implicit data linkages are broken, because the parsing process reassigns DDIDs to each individual observation. This effectively removes the real (explicit) *and* contextual (implicit) identifiers from the dataset, and thus eliminates the risk presented by such equivalence classes, while also masking unicity in the dataset. Specifically, it is not clear, without accessing the identification maps, whether the high blood pressure (DDID=0847) is assigned to a female person (DDID=9999). Furthermore, we cannot use that context to link data together to identify with certainty that DDID=0847 is *Alice's* blood pressure reading, as compared to DDID=6878. In this dataset with 6 individuals, we now know that the risk of re-identification is $r = 1/n = 1/6$ (random) and since $1/6 = 0.17 < \rho = 0.20$ we define this as an acceptable risk. Put more strongly, any dataset with at least six observations would be compliant using the Anonos JITI / DDO method. However, even if we uniquely found a single individual with high blood pressure (as is the case above), doing so does not improve our ability to link that to any individual nor to access other information using that knowledge.

Dynamism and uncertainty

While this dataset is currently seen as time-invariant (the most conservative case for analysis, and likely when a third party gains access to a full dataset only once, perhaps through capture of a transmission); this may not actually be the case when using Anonos' JITI / DDO method over time. Specifically, because Anonos does not use temporally-static identifiers, downloading the same dataset a second time could easily lead us to reasonably, but incorrectly, conclude that the new dataset references new data because the new DDIDs are dynamic and thus new identifiers reference new data *which is also reordered*. Thus, it may be possible that the second time data are transmitted, the sample size no longer seems to reference only six people, but instead might be seen as incorporating different data and thus reference a sample that may be

as large as 12. Doing so would effectively reduce the risk of re-identification so that $1/6 \leq r \leq 1/12$ (taking a simple average, $r \cong 1/8$).

Deceptive replication

Similarly, you could mask the data in this table by adding in random information (with requisite DDIDs) and similarly mask the sample size and unicity in the data. These types of deceptive replication efforts may be further compounded because new data are randomly sorted and may incorporate more newly integrated observations. If this is the case, the duplicated or incorrect observations may *give the appearance of a larger sample size*, reducing the risk to a range ($1/6 \leq r \leq 1/12$), so that, on average assuming a uniform or normal risk distribution within that range, $r = 1/8$, a reduced risk of re-identification. However, this reduction depends on the assumption that the dataset as a whole does not contain any unique information that is not anonymized (such as the IP address above), which would be obvious once replicated and thus retain the more conservative $r = 1/n$ level of risk.

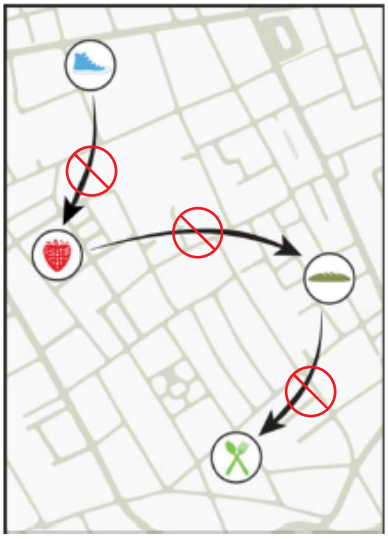
A secondary, and robust, gain from Anonos' JITI / DDO method is that we no longer know whether the two women referenced in the table above are two women or the same person measured twice. ***Challenging these assumptions can be uniquely beneficial because it forces us to question whether our basic assumptions about equivalence classes are correct, and further highlights the role of data that changes over time.*** Specifically, while we may note that here we have two people with different blood pressures, having two women the same age actually requires us to either assume that each outcome references different people (with the level of risk noted above) or to wrongly posit that these two observations reference the same person and either their health status changed over time (forcing us to question the nature of re-identification itself, since we can no longer determine when a person had the health outcome), or more likely

assuming (incorrectly again) that the woman aged 45 did not have high blood pressure because there is only one observation of high blood pressure but two observations of that person.

ddid	price
G!022	$97.30
015M	$4.33
15166	$43.78
11A	$35.81
^99M	$15.13
0099S	$12.29
56KHJ	$3.66

Derived from:
Science 30 January 2015:
Vol. 347 no.6221 pp.536-539
DOI: 10.1126/science.1256297
www.sciencemag.org/content/347/6221/536

Fig. 1 Anonos Just-In-Time-Identity (JITI) enables dynamic protection at the data element level.
The universal "no symbols" highlight that dynamically obscuring data linkages that could be aggregated by parsing recognizable static "anonymous" identifiers breaks the assumptions necessary for re-identification.



Note: Figure 1 above shows that DDIDs 6730G, SiMX, Z3O and E2Grs may be used to refer to the same user at various times. And further provides DDIDs for prices, which have been randomly ordered. This obscures data linkages that could otherwise be aggregated by parsing recognizable static "anonymous" identifiers like the identifier "7abc1a23" that was used to refer to "Scott" for each transaction in de Montjoye et al. (2015).

In either case, these reasonable assumptions force us to make incorrect conclusions, making longitudinal data *less useful to re-identification than cross-sectional data*. Specifically, the risk of re-identification is no longer inversely proportional to the number of people, but instead to the number of observations: $r \leq 1/6$. This challenges the assumptions made by de Montjoye et al. (2015), who used longitudinal and specific data linkages to uniquely identify

individuals by, as is noted in Figure 1 above provided by Anonos, breaking the assumptions required for unicity. Specifically, Figure 1 notes that while multiple data-points can be used to uniquely differentiate individuals, that the Anonos method breaks the implicit and explicit linkages between data points, and thus effectively removes the ability to represent data in this way. Specifically, we may know that an individual shopped at shop A, but not how much they paid at that shop nor which store they shopped at next (if any). ***Put conservatively, Anonos' method does not preclude the possibility that data are unique in a dataset (for example, the high blood pressure reading is unique in Tables 1 and 2 above), but makes that information useless in determining anything else about those data. It prevents unique data from being used to attempt data fusion.***

Transmission

By assigning to each person a DDID, Anonos replaces individual identities with temporally dynamic random information, replacing usual static re-identification protocols by requiring that we rely on assumptions in re-identifying. The ability to re-identify individuals by intercepting transmitted information is predicated on data linkages – various quasi-identifiers or variables that can be aggregated by parsing the identifiers. During transmission, these data are described by Anonos as being dynamically obscured via the DDID process, and transmitted as the data available above. This corresponds to the following data stream being sent from table 2 above: DDID=0967; High BP=No; DDID=3030; Lat.=40.13; DDID=4958; Age=22; etc.

If Anonos dynamic obscurity fails, the risk that this blood pressure indicates *Alice's* blood pressure is $r = 1/n$. However, the calculated risk is variable and, though dependent on sample size, the risk remains unknown to those interested in re-identification (because the total sample size is unknown). Moreover, deferring the transmission of this information and sending it

separately increases uncertainty about the context in which the data is being delivered; because data are being de-identified, we cannot be assured, without making assumptions about this equivalency, even if only longitudinal data referencing a single person's data were being delivered over a period of time, that this data referenced multiple observations of a single person who moved around between different locations, rather than of multiple people living near each other. For example, while the above represents six people because each category provides different locations and IP addresses for a wearable dynamic, if we replaced the table above with a table referencing two women aged 45 followed up for three time points, the new data would *appear* identical to data referencing one woman aged 45 but followed up for six time points. This would be especially confusing if this person moved around between areas. Again, however, given that we know where a person was at time 1, we cannot use that information to derive information about her health or location at that time. As long as the total number of application users equals or exceeds five, and no assumptions can be made about the number and types of information available during a particular transmission (i.e., we cannot know that only one person's information is being transmitted), the risk of re-identification remains acceptable even during transmission.

Let us assume the worst-case scenario: that such a transmission were caught and *revealed in its entirety*. Then we are left with the case, explicated above, where the number of users is known. Thus, $r = \frac{1}{n} = \frac{1}{6} = 0.17$, which we deem to be an acceptable risk (because $0.17 \leq 0.20$). However, if the transmission is not entirely caught (for example, if blood pressure is not caught or sample sizes differ between observed variables), then the risk of re-identification must be derived from information known about the number of users of this particular blood-pressure application at this particular site. Because the number of users must be at least six (since there

are six in our known dataset), we know that the risk of re-identification becomes bounded by the dataset and user-base specifics, and is thus at most 1/6 and at least 1/n, and thus we would say that the risk of re-identification is $r \leq 1/6$.

In this case, transmission effectively replicates the "sampling" method detailed by Matthews and Harel (2011) above; a common de-identification technique in itself. Formally, this means that Anonos JITI / DDO efforts serve to increase n by adding to it an unknown amount (k), where $k \in \mathbb{Z}$. With the addition of k, re-identification then relies on the probability $1/n = 1/(n' + k)$, where k is *unobserved*. Notably, k could easily be made up of data that is similar to the real data, but is fake, or by replicating randomly sorted data that is not differentiable from its copied counterpart. As a result, the risk decreases rapidly by the inverse of the total number of observed and *unobserved* users (n). More concretely, if we know that at *Alice's* place of work that there are 20 users of the particular application then the risk = 1/20 = 0.05, which is less than 0.20. If, however, all employees (say 350) have been provided access to the application, then the risk = $1/N_{employees} = 1/350 = 0.0026 < 0.20$. In either case, the risk is acceptable as long as the number of users in the full, accessible, dataset does not allow for $r = 1/n \geq 0.20$ (i.e, $n \geq 5$).

Optimization and efficiency

Regulatory compliance specifically requires that PD are subject to a reasonably low risk of re-identification. We have shown above that the JITI / DDO method can reduce that risk. However, it may be inefficient to dynamically identify every piece of information at all times, so understanding the necessity of such levels of security to maintaining JITI / DDO compliancy may be useful. Above, we suggested that we could parse data by cell into randomized data with unique DDIDs. However, we could maintain many of the same levels of security by viewing the main dataset as a matrix of matrices (i.e., that each matrix $\boldsymbol{H}$ contains '$\boldsymbol{j}$' matrices within which

the data reside). As such, DDIDs could be used to secure data not by providing DDIDs to each cell in the dataset, but instead to each matrix of cells in the dataset. This would provide much of the same level of security discussed above, but would be much more computationally efficient.

Specifically, modifying Table 1 above we provide the following dataset as a set of groups that are defined by the DDID to create Table 3 below, where each level of gray (of which there are j=6 made up of 3 rows and between 1 and 2 columns of data) signifies a different dataset formed of sequential or random pieces of information that could only be reassembled, like a puzzle, using the key. Here, the blocks were predefined to overlap with each type of variable in part because this overlap is easier to see and manage, but is also in many ways more secure.

Table 3. Hypothetical dataset (***H***) collected from multiple smartphones on the same network by a blood pressure application in December 2014

IP address	Latitude	Longitude	Age	Sex	High blood pressure
192.168.0.0	40.13	-79.85	32	M	No
192.168.101.201	40.13	-79.86	45	F	Yes
192.168.4.57	40.15	-79.54	39	M	No
192.168.75.122	40.10	-79.59	45	F	No
192.168.22.40	40.29	-79.54	56	M	No
192.168.1.220	40.29	-79.56	22	F	No

It would be evident if one happened to receive one particular sub-matrix, that there was a respondent with high blood pressure (the darkest gray). However, it would be impossible to ensure that this respondent was our fictional individual, “Alice” as above. Nevertheless, it would be entirely feasible to know this if certain types of data were contained within that dataset, and

thus security would only be ensured if types of data were contained separately from each other, and if the matrix mapping were not unique in itself (i.e., if matrix ***H*** could be reasonably made of by assembling these ***j*** pieces in a number of ways). Here we noted that data could be differentiated in this way: the IP address column is white because we assume it would be deleted, while the data in the blood pressure chart are held in pieces that could be made up of information in *n* (here 6) ways. As such, this provides similar security as does the method above with one caveat: if data are longitudinal and variables are stored in concert, and the outcome are sufficiently specific, then there is a small chance of matching data types together. Nevertheless, this would be reasonably solved by implementing variation in the levels of security promoted by the types of data so that publicly available data are stored in *j* cells while more sensitive data are stored in single cells without linked data. In this example, let us suggest that, under the worst-case scenario, we received the data in full but separated by shade of gray into six datasets. In this scenario, we would know because of our mapping only that one respondent had high blood pressure, resulting in a risk of re-identification of 1/6, which we have defined as acceptable. However, if this were not the worst-case scenario and only a subset of the data were received (the darkest gray box, for example) then the risk of re-identification is *at most* 1/6 and is at least 1/*n* where *n* could include the entire base of potential users.

As noted above, this would be further secured by the occlusion tactics described above, and could be modified to secure other types of information than health data, subject to risk analysis about the specificity of that data. This application of this type of analysis has two added notes regarding specific data. Specific data could be considered to be largely "identifiable" information and would need to be separated from other forms of information to maintain privacy. Finally, longitudinal data (a form of specific data) can be stored in two ways: wide, with each

variable noting an observation; or long, with each observation identified within the implicit structure underlying a dataset (as in Figure 1 above). In either case, this method could be made robust to long or wide data depending on mapping techniques and, if necessary, random sorting. Crucially, in this scenario PD would still 1) not be subject to data fusion and 2) be kept private even if unicity were achieved in the data itself.

Differential levels of security

One benefit of this type of optimization in conjunction with dynamic capabilities is that it facilitates the ability for users to manage security flexibly. Specifically, if some data were considered a greater risk than other data, it would be possible to vary the level of security used to secure different types of data and to secure data for different purveyors. For example, let us imagine that age and sex needed less security levels than a medical diagnosis of schizophrenia. It would be possible to differentiate them and use different mechanisms to organize them. Such variation matches how comfortable individuals might feel sharing information in person. For example, one could keep the age and sex variables integrated but randomly sorted. The diagnoses could, on the other hand, be differentiated and dynamically de-identified and excluded from the dataset's user base unless specifically provided by the security policy. This differentiation would both eradicate the risk of data fusion and would minimize the risk of re-identification. However, it could allow easier access to basic demographic information for accepted purposes. In this way, users could, with the guidance of policymakers *or with the explicit permission of individuals from whom they have collected data,* readily provide a sliding scale of coverage where different types of data are differentially secure. ***In practice, this would imply that Anonos could implement what could be termed a "programmatic policy", or a digitized representation of***

policy decisions that defines,* a priori*, how data are shared by specifying which data are shared when and with whom.

Unknown Third Parties

The above analyses are sufficient to ensure de-identification in the traditional, static, sense. However, we live in an increasingly demanding and dynamic world, with increasingly opaque privacy protocols. We may therefore reasonably assume that the end-user is not yet defined and that more complex associations may arise. We also may encounter the real outcome that an end-user may try to re-identify individuals in their own data surreptitiously without the knowledge of Anonos. We may thus be interested in knowing whether an unknown third party (U3P), not bound by data privacy standards and in possession of substantial resources (human, financial, or political), could *surreptitiously* manipulate the data to facilitate re-identification. If this is the case, then interested parties might have strong incentives to try to *find or create* a circumstance where re-identification could be made easier. These third parties may be internationally based and thus not easily dis-incentivized by standard legal considerations.

To examine this possibility, we asked the following hypothetical question: *could an end-user, with a previously specified user base, ask specific questions in order to facilitate re-identification?* Put more specifically, in an attempt to identify a target user, could a U3P modify an existing membership's data collection routine, containing the targeted user, to modify their data collection routine (but not their user base or Graphical User Interface / GUI) to clandestinely determine that user while incurring an unacceptable level of risk ($\rho > 0.20$) that health data refer to a particular individual (for ease, *Alice*)?

The most readily available technique is to add questions or indicators that would easily facilitate such re-identification. For example, the risk of re-identification could be increased by

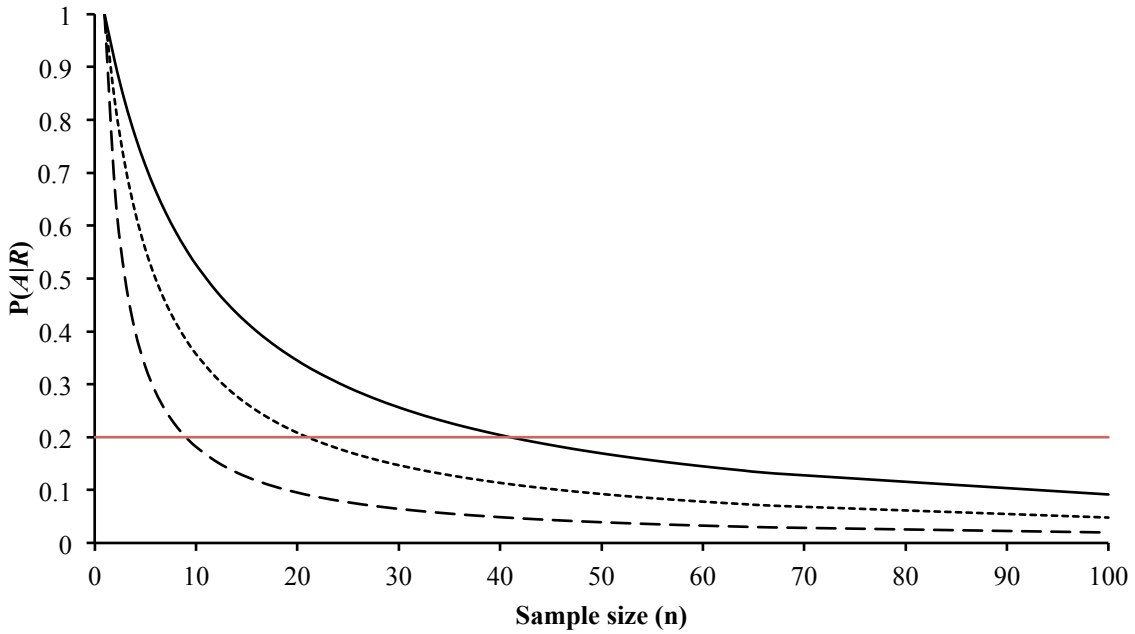
defining multiple non-threatening questions that overlap in order to increase unicity and facilitate the unique identification a particular person, or by linking smartphone data or metadata (including, for example, GPS information) to publicly available information. However, because identifiable information and characteristic indicators, which might be easily added to the application in order to expressly identify the individual of interest (i.e., to maximize the risk of re-identification), are subject to Anonos' JITI / DDO method, these linkages are readily dealt with as noted above. We must therefore assume the U3P could simply access the full dataset with the identifiers from an authorized user; thus, the worst-case scenario is that they would know a person was part of the data collected. It may be possible then to gain the full dataset, but using Anonos' method, these data will not be linked and thus the result will not be more informative than that – you would know that a person was part of the data, and that there is a $1/n$ risk that any indicator, including high blood pressure, a relatively low probability. Because these data are not linked, we know that asking identifiable or characteristic questions could only be used to determine the health of a particular individual with a risk of re-identification of $1/n$.

If identifiable or characteristic data are not useful, it may still be possible to determine/create a situation in which information is both 1) interesting in its own right, and 2) sufficiently specific to determine with risk ($r > \rho = 0.20$) that a person fits the outcome suggested. This quest is trivial if the person of interest does not, or is not known to, use the application during the period of time under examination, since the risk will then always be $0/n = 0$. However, in our hypothetical situation, the U3P would know that the user's information was contained within the dataset provided. Thus, the data requested must, in a single unlinked variable, reference an outcome where its specificity and risk could be sufficient to identify an individual's information solely on its specifics. This is easiest when the potential outcome is

strange rare (either the disease or lack of disease) since the risk of identification relies on assumptions and unicity in that dataset.

To maximize re-identification risks, we must therefore create *specific* data that are both the health variable of interest and sufficiently unique to successfully identify the information desired. This is a tall order and highly unlikely in any normal dataset, so an interloper asking these types of questions might be obvious to the respondents. In such a situation, we might reasonably assume most individuals to be free of the disease, and that we have reason to believe that the risk that *Alice* has the disease is M times greater than the normal population. Nevertheless, we want then to know what the likelihood is that *Alice* (A) has condition R, given that R is observed in the dataset (denoted, $P(A|R)$). This calculation can be solved using Bayes' Theorem. The probability *Alice* has the disease is: $P(A|R) = P(R|A)*P(A)/P(R)$. These other probabilities are either known or can be guessed. For example, the probability that any observation is *Alice's* is $P(A) = 1/n$. The probability that any sample contains an observation of that is $P(R \text{ particular disease}) = (R*(n-1)+M*R)/n = R*(n-1+M)/n$, where R (such that $0 \leq R \leq 1$) is the risk of the disease. We believe, from our expectations derived from external observation, that if *Alice* has a risk M times the normal risk (R) of observing the outcome such that $Q = 1-R$, then the probability of a positive outcome given that *Alice* is in the sample is $P(R|A) = MR$. Thus, we have from Bayes' Theorem that $P(A|R) = P(R|A)*P(A)/P(R) = (MR*1/n)/(R*(n-1+M)/n) = M/(n-1+M)$.

Figure 2. Likelihood of re-identification under the hypothetical condition that data are manipulated in order to engineer the best conditions possible to identify individuals, with an estimated *M* of 2 (long dashes), 5 (dotted), and 10 (solid) by sample size.



Simulations estimating the risk of re-identification given a single positive observation follows Figure 2 above. We have here assumed a range of relative risks ranging from conservative ($M = 2$) to medium ($M = 5$) to very large ($M = 10$). This range of relative risks (M) was allowed to range from 2-to-10 to reflect the range often seen for predictors in epidemiological research, and because at most M references the difference between a risk of 2% (a rare outcome) and 20% (the risk necessary to be reasonably certain $\rho = 0.20$). Notably, risks much higher than 2% become decreasingly likely to enable an $M = 10$ outcome because when the population's risk approaches 10%, the personal risk must approach 100% (i.e., 10*10%), a known certainty, and thus the need for re-identification is unnecessary to begin with.

Figure 2 provides a more conservative estimate of the risk of re-identification than traditional methods. This estimate suggests that in the worst possible situations that Anonos' method is robust to intentional privacy intrusions by a U3P undertaken with express knowledge of the Anonos method, as long as total sample size exceeds 41 individuals (the point where the solid black line ($M = 10$) crosses $\rho = 0.20$). Notably, while it is unlikely that all data are going to need this level of privacy, it is reasonable to suggest that when data are treated in this manner that they achieve or surpass this stringent level of security.

Discussion

In this analysis, we described Anonos' JITI / DDO method as starting with the premise of blending existing methods of de-identification, including for example sampling, suppression and the potential addition of noise, with novel temporally dynamic identifiers and data parsing protocols. We have analyzed the risk of re-identification, finding that the Anonos JITI / DDO method can drastically reduce the risk of re-identification, even for specific data. Moreover, these analyses we found that, using the Anonos JITI / DDO method, data were kept private during both transmission and storage, even from the application developers. Specifically, we found that re-identification risks were minimized and could be reduced under the generally accepted statistical and scientific principles and methods for rendering information not individually identifiable (threshold value (here defined as $\rho = 0.20$)), when the total sample size exceeded five analytic units (e.g., individuals, households, online identities, IP addresses, etc.). Moreover, we discussed the potential for JITI / DDO processes to be applied to blocks of data rather than individual observations. We further found that the level of security could be managed by variable differentiation and de-linkage, so that some information, such as basic demographic information was not de-identified but other information was at the same time subjected to the

DDO process. We further discussed the potential for both policymakers and for individuals from whom the data are collected to help define the level of security of particular data.

Risk of re-identification

The risk of re-identification is limited by constraining assumptions that can be made about data contents and structure (Kifer & Machanavajjhala, 2011). *Anonos works by breaking the assumptions that are encoded in datasets and used by others to achieve re-identification.* Breaking these assumptions has a number of benefits, but the most important one is that it makes the both re-identification and the risk of re-identification difficult to ascertain with any level of certainty without further gaining access to the complete, unadulterated, dataset. Anonos JITI / DDO does this in a few main ways discussed below.

Being dynamic helps. DDIDs provide a level of protection from data and the misuse of data that are not available now. For example, DDIDs necessarily re-integrate randomized follow-up information from the same individuals if data were downloaded later, and thus serve to increase sample size and reduce re-identification risks while reducing our ability to make assumptions about the completeness of the data. Secondly, the data differentiate instances from one another, making assumptions about the completeness of data, and their reference population, less clear. Third, Anonos JITI / DDO efforts work well during transmission to effectively occlude shared information and to maintain security even with characteristic and specific data. Finally, the method can be made robust even to those who are engaged in collecting the data, making data privacy clear and enforcing data use agreements even when unknown third parties are engaged in using the data.

Flexibility of privacy and security

This technology enforced decoding of DDIDs could apply broadly, within a single

deployment of Anonos, but it would be possible to overlap multiple cascading rule sets, with an agreed-upon hierarchical relationship, to govern usage of any given primary data table. In practice, this could mean that a lead country's Data Protection Authority (DPA) might define the highest-ranking set of PD access rules, but another concerned party might also insert its own set of PD access rules that may be more stringent. These rules might be applied differently to different types of data within the same dataset. In this event, Anonos can be configured to ensure that no PD access is possible unless both cascaded sets of DPA access rules are enforced when the query is made. Conversely, JITI / DDO could provide flexible controls necessary to support hierarchical handling of various data privacy requirements.

Programmatic policy

The ability to deliver on the many promises of big data in linking together individuals with institutions, clinicians, or researchers, for example, is predicated on this ability to support differing privacy requirements depending on the nature and source of data. Anonos provides a way to automatically and digitally enforce such privacy policies. For example, consumer health data collected using electronic health records, mobile health applications, and social networking sites may be accessed and data may be useful and available. At the same time, financial data may be transcribed into the same data using the same devices. However, PD is at the same time regulated by privacy and security requirements under a given country's privacy and health privacy acts and may further be subject to specific privacy policies and terms and conditions depending on user preferences for specific websites, devices and applications. The JITI key itself encodes both the rules necessary to recover the source value from at least one DDID and flexible programmatic policies, or a digitized representation of the privacy policy that is subject to observation, enforcement, and audit. Therefore, if necessary rules and constraints are not being

observed and enforced, either because there is 1) a mismatch between a user's permissions and the query that user is trying to submit or 2) access was once granted but has since been revoked or expired, then no DDIDs may be decoded.

Conclusion

The Anonos JITI / DDO invention and protocol mitigates the risk of re-identification by repudiating assumptions about explicit and implicit data linkages. It can therefore ensure privacy even when the dataset as a whole contains characteristic or specific data, such as when single individuals are followed over time or specific details such as geographic location are observed. The flexibility of the JITI key mechanism ensures that the Anonos Policy-Driven Data Management Platform, even in cases where a single data element of PD must be protected by a single JITI key, programmatically enforces granular rules for access. We also found that, even when individuals worked to design a situation favoring re-identification, Anonos' method continued to minimize the risk of re-identification by first removing the risk of characteristic re-identification, while repudiating the ability to make assumptions about the structure of the data, and also by limiting the risk of specific re-identification to acceptable levels given sample size limitations. We then identified opportunities to both: further occlude data in cases of small numbers of observations, and optimize occlusion to facilitate large-scale data management. ***It is the author's opinion from the analyses conducted and described herein that, subject to oversight and sample size limitations, Anonos' JITI / DDO method substantially mitigates to a statistically acceptable level the ability to single out, infer about, or link data to an individual so that personal data remains private.***

Author Biosketch

Sean Clouston, PhD, is a *Fulbright* scholar who earned his Bachelor's in Arts in Mathematics and Sociology. He did his doctoral work at *McGill University* in statistics, epidemiology, and demography and is currently employed as Core Faculty in the Program in Public Health, and an Assistant Professor of Preventive Medicine at *Stony Brook University*, part of the State University of New York. Dr. Clouston uses quantitative analysis on high dimensional data to examine questions relating to the distribution and determinants of disease both in the United States and globally.

References

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association, 17*(2), 169-177.

Berinato, S. (2015, February 9th). There's no such thing as anonymous data. *Harvard Business Review*.

Castro, D. (2015). FTC's Internet of Things Report Misses the Mark [Press release]

Chen, D., & Zhao, H. (2012). *Data security and privacy protection issues in cloud computing.* Paper presented at the Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on.

Cox, L. (1994). Matrix masking methods for disclosure limitation in microdata. *Survey methodology, 20*(2), 165-169.

de Montjoye, Y.-A., Radaelli, L., Singh, V. K., & Pentland, A. S. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science, 347*(6221), 536-539. doi:10.1126/science.1256297

El Emam, K. (2015). Is it safe to anonymize data?

El Emam, K., Dankar, F. K., Vaillancourt, R., Roffey, T., & Lysyk, M. (2009). Evaluating the risk of re-identification of patients from hospital prescription records. *The Canadian journal of hospital pharmacy, 62*(4), 307.

Federal Trade Commission. (2015). *Internet of things: Privacy & security in a connected world*. Washington, DC: Federal Trade Commission.

Herz, J. (2014). Wearables are totally failing the people who need them most. *Wired*.

Hotz, R. L. (2015, January 29). Metadata Can Expose Person's Identity Even When Name Isn't; Researchers Use New Analytic Formula. *Wall Street Journal*.

Kifer, D., & Machanavajjhala, A. (2011). *No free lunch in data privacy.* Paper presented at the Proceedings of the 2011 ACM SIGMOD International Conference on Management of data.

Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys, 5*, 1-29.

Measure DHS, & ICF International. (2013). *Demographic and Health Surveys*. MD5.

Rivera, J., & van der Meulen, R. (2014). Gartner says the Internet of Things will traansform the data center. *Gartner*. Retrieved from http://www.gartner.com/newsroom/id/2684616

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics, 9*(2), 461-468.

Tyrrell, C. (2014). Countering HITECH privacy risks fro internet of things products. *HITECH Answers*.

Warren, N. (2014). Dynamic Data Obscurity. *Category Archives.*