

A New Model for Privacy Measurement in the Database Age

Brad Rosen

May 5, 2003

Abstract

Formulation of a new model to quantitatively measure expected privacy. Summarization of the status quo and a brief experiment tracking the flow of information in website databases.

Keywords

database security privacy bit-descriptor trusted-entity domains SPAM

1 Introduction

*It is the responsibility of the sender to make sure
the receiver understands the message.*

-Joseph Batten

The term *Information Age*, coined by Alvin Toffler in the 1970's, has been used for the past thirty years as a hallmark of the ubiquitous nature of computers and their ability to aggregate data for us. Yet, the mere collection of data, of information, serves very little useful purpose. It is the organized storage, easy retrieval, and cogent presentation of data that has had a noticeable effect. The word *database* comes to mind quickly.

It is not, then, the *Information* that is paramount, but rather the methods used to store and retrieve that information. In effect, this chalks up most, if not all of the "Information Age" hype to databases. The modern hierarchical file system, organized (at least in theory) to allow for quick, mnemonic access to important documents, is a database in some sense of the term.¹ In the confines of this paper, a *database* will refer to any "large" electronic repository of data.² It is sufficient to say that databases have enabled an all-pervasive information-oriented society, in which we can easily

¹At the current time, Microsoft has announced that it is endeavoring to replace the currently used hierarchical filesystem with a sql-based filesystem. Simultaneously, Apple may be implementing their "Piles" filesystem concept, patented in 1994. The Piles system allows content-addressable grouping and querying in what has been described as a "databaseque" fashion.

²Justification: Any such datastore will inevitably be represented using a modern database package; one would be hard put to analyze large datasets using merely flat files.

collect, store, aggregate, retrieve and analyze **virtually any salient event**. For now, consider *privacy* to be the ability of entities involved in those events to decide if non-involved entities may know of those events, and *security* as the techniques used to ensure that the privacy wishes of the former entities are not violated.

Like much of the nascent Information Technology industry, springing as it did from well-meaning Computer Scientists, early database theorists and implementors were not yet in a mindset to think about the implications of having easy access to all that stored data. (Rather, they were busy hard-coding disk sector numbers because no one had come up with a decent disk abstraction yet...) Concepts such as *Security* and *Privacy* were ill-defined at best. Corporations and Universities wishing to utilize databases were far busier worrying about the staggeringly prohibitive cost of hardware and the difficulty of finding the (relatively) small number of programmers with sufficient technical knowledge than they were about philosophizing on the societal impact of data collection.

Unfortunately, due to advances in database technology, namely the introduction of relational databases, and the rapid advances in electronics, the power of databases to store and aggregate information grew far faster than the public awareness of that power. As such, oftentimes privacy was violated without the intent to do so on the collection side or the realization that it was being done on the subject side. Disparate data, when grouped together, often yield shocking results. Imagine that your name is Ebenezer Quincy Snodgrass III. [7] Now imagine a convention center full of people, merchants, bankers, lifeguards, babysitters, deli owners, judges, free-clinic physicians, and prostitutes. If they all had but one piece of information about Ebenezer Quincy Snodgrass III, or Ebenezer Q. Snodgrass III, or even just Ebenzer Snodgrass, imagine the massive amount of information they would have about your life, personal habits, vices and hobbies. This is nothing more than a simple database-join on the key "Ebenezer (Quincy) Snodgrass."

How much easier is it to track someone if their name is Ebenezer Snodgrass than if their name is John Smith? What actions or habits can a person have that will contribute to the wealth of data about them, mere database-joins away from aggregation? The goal of this paper is to provide the framework for a new model of answering these types of questions in a quantitative manner.

Within the scope of this paper, we will examine the statue quo of databases and their affects on privacy, see the results of a brief experiment designed to give a quantitative index for measuring that privacy, and see how this new model incorporates these features.

Due to time constraints imposed upon this project, it is not intended to be a complete model but rather groundwork on which further research can draw. A number of avenues will be partially discussed but not fleshed out in full, leaving room for further work.

2 Status Quo

2.1 Overview

In the status quo, the raw processing power of databases has led to massive aggregation of data. The relatively lax, if existent, governmental controls have allowed for widespread abuse in a mostly self-regulated commercial industry. By dividing privacy into three specific domains, we lay a roadmap for developing our privacy model.

2.2 Domains

When trying to discuss the current state of Privacy with respect to databases, it helps to split the discussion into three distinct, but possibly overlapping, domains. The consumer domain, the private individual domain, and the governmental domain. For the scope of this paper, we will be concerned almost exclusively with the governmental and consumer domains, and some of the places where they overlap with the private individual domain. The description of the private individual domain is included here for completeness.

2.2.1 Consumers

The consumer domain is defined by the interaction of smaller entities (e.g. people) purchasing goods and/or services from larger entities (e.g. corporations). This is the primary domain of opt-in/opt-out mailing lists, junk mail, direct marketing, tele-marketers, and SPAM. A large portion of what is currently labeled “privacy” falls in the scope of this domain. Databases in the consumer domain typically refer to firm’s customer databases and mailing lists.

2.2.2 Individuals

Individual privacy is defined as the realm of private-life interactions. With respect to American individuals, the main focus of this paper, this realm includes all of the liberties reserved by the Constitution and the Bill of Rights. This realm interacts heavily with the governmental domain, and to a lesser degree, the consumer domain. One of the major threats in the individual domain is the potential to be (incorrectly) labeled as a criminal, sex offender, terrorist, or other such persona non grata. For the scope of this paper, we will not be concerned so much with this domain directly, but rather how misuse of data from other domains may bleed into it. Databases in the individual domain often intersect with the governmental domain, but also include files maintained by private detectives, stalkers, hospitals and other health institutions.

2.2.3 Government

The governmental domain is the body of knowledge necessary for efficient operation of federal, state and local governments or the datastores mandated by law that these organizations maintain. Databases in the governmental domain cover tax data, motor vehicle and firearm registrations, sex offender databases, public records, court proceedings, and other such collected data. As of late, this sector has gotten substantial press, mostly due to anti-terrorism efforts and President Bush's plan to create a unified national database.

2.3 Domain Interactions

Of particular interest is the flow of data between these domains. The Organization for Economic Co-operation and Development (OECD), an "international organisation helping governments tackle the economic, social and governance challenges of a globalised economy," has a comprehensive list of guidelines for "Transborder Data Flow." [9] These principles are meant to be applied to legally mandated data collection, yet it is useful here to apply them to all aspects of domestic data collection. The United States is a member of the OECD, and as such, should comply with their recommendations (in theory). The OECD has eight salient principles, most of which are fairly non-technical. For the purposes of this paper, we shall restrict ourselves to four principles and refine their specificity. These principles are:

- **Collection Limitation:** There should be limits to the collection of personal data; collected data should be obtained by lawful means with the knowledge or consent of the subject.
- **Data Quality:** Data must be fully up to date and accurate.
- **Use Limitation:** Without express consent, data may not be made available or used except for the purpose it was collected for.
- **Individual Participation:** Individuals must be able to know what data is being collected about them, and be able to view that data. If for reasons they cannot view the data, they must be able to appeal the reasoning. They must be able to appeal any incorrect data and have that data fixed or expunged.

2.3.1 Private-Consumer

The private-consumer interaction is troublesome, yet would be fairly easy to regulate. Services such as Equifax that have traditionally sold consumer data that may be strongly linked to private data, such as online purchasing habits. Oftentimes, data is collected with sweeping reasons that are far too vague. Typically, far more data than necessary is required on certain forms, violating the Collection Limitation. Furthermore, these consumer databases assume ownership of the data, violating the Use Limitation by reserving themselves the right to do whatever they please with that data.

The New York City Metropolitan Transit Authority (MTA) maintains a commercial MetroCard system for city subways. A central database records the time and location when a MetroCard is swiped. Anyone mailing in a MetroCard is given, cost free, complete access to the list of stations where that card has been used. According to Garfinkle, “all anyone has to do is filch a card, and all of the card’s records are open for the asking.” [5] The implications for personal privacy are quite shocking. For example, a celebrity or politician frequenting “red light” or “homosexual” districts, regardless of their personal discretion, could be terminally damaged by the publication of their MetroCard record. In going perhaps too far to comply with the Participation Principle, the Use Principle has been violated. Furthermore, many commuters may not know that their movements are being tracked, violating the Collection Limitation.

This case of domain interaction will play an important role in the synthesis of our privacy model.

2.3.2 Private-Government

Under the scope of this paper, the interaction between private and government domains is not relevant except in the transitive case where data flows from the consumer domain to the government domain and then affects individual privacy. Typically, private-government interactions violate all four of the principals outlined in Section 2.3.

2.3.3 Consumer - Government

As Feigenbaum suggests, this is the most dangerous interaction due to the nature of the forms that many consumers fill out, and that “People will write anything for a free coupon.” Furthermore, the FBI has recently relaxed the standards to which it holds its own database accuracy.[3] The danger lies in the fact that governmental organizations may acquire this information from the consumer domain. This technique is “not just unfair but *woefully* error prone” [4] due to the crossing of domains. Feigenbaum refers to this as “using data created in low security environments in serious contexts.” [4] Compounding the situation, these types of situations may also affect individual privacy, as inaccuracies from consumer databases find their way into governmental ones. Much like private-government interactions, consumer-government interactions also often violate all four of the principles outlined in Section 2.3.

2.4 Security

To paraphrase Section 1, Security is the combination of technology and policy aimed at preventing the unauthorized destruction, modification, collection and/or use of private data. There are two distinct threats to security: foreign intrusion and inside jobs.

2.4.1 Foreign Intrusion

Movies like *Sneakers*, *Tron*, and more recently, *The Net* have popularized the notion of the hacker, working alone or in groups to penetrate sensitive systems and steal data. The simple fact of the matter is that computer systems have sufficiently matured to the point where these types of fantasies have been relegated almost exclusively to the realm of the imagination. While it is not unheard of for the occasional successful theft of credit card numbers or customer data [1], it is far more common to hear about accidental leakage of customer data. [6] [10] For home users, the threat of direct hacking into their computer systems is relatively negligible.

For businesses, enterprise-level computing has become firmly entrenched in a backup culture. With regards to data loss, firms have far more to fear from an angry ex-employee with a sledgehammer and the key to the datacenter than they do from a 12 year old with a script downloaded from IRC. Likewise, there is a greater threat of an employee selling corporate data to a direct competitor than there is of a malicious hacker breaking in, stealing that data and doing the same.

2.5 Inside Jobs

Companies are particularly at risk from employees “just about to leave the company” [8] either voluntarily, or more likely, because they have been laid off. In some cases, outgoing employees may maliciously destroy data, while in others they may try to take data which they feel has some resale value. [7] Furthermore, even employees with job security may cause data loss to fulfill a personal agenda or steal data in an attempt to further themselves financially. While data loss is a nuisance, short of burning down a building and destroying all off site backups, employees would have a hard time destroying all company data. Spot-deletions or deliberate misinformation is far harder to check for. Luong recommends logging transactions by user login and the monitoring of outgoing employees’ data accesses. [7]

In the post-Enron era, corporate accountability has become a buzzword that has significantly reduced incentives for data theft, as the resale market is no longer quite as robust. However, the critical point remains with the employees who routinely use sensitive data. An employee can always claim they called up a record by accident, having hit PrintScrn or used a digital camera to take a picture of the data. [7] In the logs, it will show that the data was accessed merely for a few seconds, indistinguishable from an actual accident. Short of taped video-monitoring of every terminal with access, actually suggested by Garfinkle to resolve the dilemma of restricting hospital record access while providing accessibility in case of need [5], there is no way of assuring that data can never fall outside the scope it was intended. (After all, who would monitor the video monitors? Even if all terminals were being watched in real time by a human, if you could bribe an employee and his or her video monitor....) At best, we can hold the party for leaking that data accountable for his or her actions.

This means that once data has left its generating source (the individual it refers to) it is no longer secure. To a certain degree, all released data has the risk of being widely accessible. This will be crucial to our privacy model.

3 Experiment

3.1 Overview

The goal of our privacy model is to be able to measure the amount of data about an entity that is available. One of the most hotly-tracked current issues is consumer privacy, specifically the sanctity of email addresses. SPAM, or unwanted and unsolicited email advertisements is a huge problem for many internet users. Billions of these messages are blocked, yet millions still get through. [2] As there are few, if any, Federal laws mandating privacy and States have been relatively slow in passing them (Virginia a notably recent exception), the industry is almost entirely self-regulatory. Many online websites post a public privacy policy that states what rights they reserve towards submitter user (consumer) data and how they will use it. They often share this data with other firms, usually referred to as valued partners, trusted partners, affiliates or other such terms.

This experiment will attempt to give a basis for website data-use tracking. By creating hard-to-guess email addresses and giving them to exactly one website, we can see exactly how much that website is sharing its customer database. Applied on a large scale, this would allow an index of quantitative numbers to be formed on a per site basis, which we will use in our privacy model. Furthermore, we will attempt to gauge if these websites honor their unsubscribe requests, as another factor in our index.

In short, we are attempting to answer the question, exactly what is xyz.com doing with its databases?

3.2 Design

The design of the experiment was fairly straightforward. Sign up for fifteen websites³ with varied target audiences and a wide range of published privacy policies. For each website, three distinct email addresses would be used, divided into three groups.

The purpose of the three groups was to gauge the effects of timing. Group one was a single hour, group two was three days, and group three was one full week. For subscription based websites, this time period represented the time period between the “subscribe” and “unsubscribe” requests. For the login-based websites, this time period represented the amount of time between account setup and the “hiding”⁴ of the email address. Finally on sites where account deletions are possible, the time period is between account creation and account deletion.

An internet host with a unique and non-public name with randomly generated email addresses was used to eliminate contamination by brute-force spamming. The host was on the local Yale network with the vanity IP address bandolero.oc.yale.edu. The email addresses were pseudo-random in that

³The following websites with www and com omitted: 1800flowers 20megsfree amazon ebay facethejury giantlinks hotcoupons matchmaker m-w paltalk slashdot smartsources talkcity valpak yahoo

⁴Websites such as facethejury and matchmaker allow you to prevent other users from seeing your real email address.

they consisted of three random alphabetic characters, an underscore, the group number, an underscore, and three more random alphanumeric characters. (e.g. `sxn_g1_h7mq@bandolero.oc.yale.edu`)

For each website, all three accounts were created in direct succession⁵ and then deactivated in whatever manner appropriate. All mail was accepted for three weeks on all forty-five accounts. An email address signed up for services at `foo.com` would expect emails from `foo.com`, but not `bar.com` or `boo.com`. “Spam” mailings were those defined as mail from an unexpected domain, or mail from an expected domain past the grace period documented in the unsubscription request. In the case where the unsubscription notification was vague about the number of days, three business days was assumed. If the unsubscription request gave no mention about time necessary to remove, the removal was considered instantaneous.

3.2.1 In Defense of Site IDs

At first, I used a script to generate randomized Site IDs so I would not be biased in my interpretation of the results. However, lacking the time to do a critical analysis of the individual privacy policies of each site, I do not feel it is fair to publish the exact linkages to Site ID. Currently, in the plans to replicate this experiment, I will do an in-depth analysis of the privacy policies of the sites involved in the study and then publish the comparison between actual results and posted policy.

3.3 Results

Table 1 shows all the initial spam counts, ordered by the randomly assigned Site IDs. Table 2 shows the Site ID’s sorted in order of the total number of spam emails received by all three addresses used on that particular site. This ordering will be the *Total-Spam ordering*, and will be used for the ordering of subsequent charts.

At first glance, the data indicates both sites with no or low numbers of spam emails and sites with high spam counts. Looking for further patterns in the data, I attempted to separate unexpected domain emails from expected domain emails. There were zero expected domain emails that violated self-imposed grace periods⁶ or the arbitrary three business day period I had assigned.

In Table 3 the initial data-set has been split into the specific groupings enumerated in the design. As expected, there is a positive correlation between time passed “in the system” and amount of spam received. There are two entries 7,8 that seem to be entirely independent of time. In Table 4, it is somewhat easier to see that there appears to be a significant jump between groups two and three in the lower half of the table.

While the data seemed significant, a brief peek at the actual mailboxes themselves revealed something not reflected in the data. The original determination for spam had only checked the “From” header against the expected domain. It did not, however, check for the content of the messages. It seemed that there were two types of unexpected domain messages. The first consisted of relatively

⁵A few services required email confirmation via reply or following a URL. These were done manually.

⁶None of which were egregious.

Table 1: Raw Data

Site ID	Total "Spam"
1	0
2	112
3	179
4	5
5	0
6	647
7	130
8	261
9	0
10	36
11	11
12	297
13	75
14	1
15	58

Table 2: Sorted by Spam Count

Site ID	Total "Spam"
5	0
9	0
1	0
14	1
4	5
11	11
10	36
15	58
13	75
2	112
7	130
3	179
8	261
12	297
6	647

Table 3: Grouped Raw Data

Site ID	G1	G2	G3
1	0	0	0
2	11	21	80
3	50	58	71
4	0	1	4
5	0	0	0
6	78	280	289
7	41	44	45
8	84	89	88
9	0	0	0
10	3	9	24
11	0	3	8
12	12	93	192
13	1	14	60
14	0	0	1
15	1	15	42

Table 4: Grouped with Total-Spam Ordering

Site ID	G1	G2	G3
5	0	0	0
9	0	0	0
1	0	0	0
14	0	0	1
4	0	1	4
11	0	3	8
10	3	9	24
15	1	15	42
13	1	14	60
2	11	21	80
7	41	44	45
3	50	58	71
8	84	89	88
12	12	93	192
6	78	280	289

well worded, coherent marketing pitches for legitimate products with accurate subject lines and an air of legitimacy, while the others were sensationalistic, with garbage characters⁷ and nonsensical subject lines, often advertising pornography or other taboo goods and illicit substances.

Table 5 shows the grouping by Reply-To and From headers⁸. If the domains of the From and Reply-To were the same⁹ then the email was grouped into GXEQ; non-matching domains were counted in GXNEQ.

Whereas previously, sites 7 and 8 had appeared fairly similar in their operation, this new grouping illustrates the differences between the two. Furthermore, most sites in the higher spam count region were typically domination by one spam type or the other.

As a side note, an errant call to ls drew my attention to the fact that three files were *very* small. A quick perusal of those inboxes revealed that the site in question had not sent a single message.

Table 5: Final Results (Sorted And Grouped)

Site ID	G1EQ	G1NEQ	G2EQ	G2NEQ	G3EQ	G3NEQ
5	0	0	0	0	0	0
9	0	0	0	0	0	0
1	0	0	0	0	0	0
14	0	0	0	0	1	0
4	0	0	1	0	4	0
11	0	0	3	0	8	0
10	1	2	6	3	19	5
15	1	0	15	0	38	4
13	0	1	10	4	47	13
2	2	9	6	15	21	59
7	38	3	41	3	42	3
3	6	44	11	47	23	48
8	1	83	4	85	7	81
12	4	8	73	20	118	74
6	10	68	26	254	29	260

3.4 Discussion, Conclusions and Implications

It is fairly reassuring to know that there were zero expected domain emails that violated the grace period. This means that all of these sites to actually allow an individual to terminate their membership.

Some of the large jumps between the three day group and the seven day group (Table 5) suggests the outflow of data from these sources is happening on a weekly basis in lieu of per diem or continuous

⁷Presumably to bypass spam-filtering products.

⁸Messages lacking an explicit Reply-To header were considered part of the "equal" category

⁹Allowing for From: mary@foo.com but Reply-To: marys.secretary_bob@foo.com

basis.

The profound distinction between the seemingly similar sites 7 and 8 raises an important point. There seem to be two distinct modes of operation, one in which websites specifically target legitimate marketing organizations, the other in which customer data is sold to the highest bidder, regardless of his or her intent. Site 7 had a much higher incidence of semi-legitimate messages, where site 8 was almost exclusively a source of exceedingly offensive material. While it is entirely possible that site 7 merely had a miscommunication with the partners it has shared data with and that the flow of email will eventually stop, it is doubtful such is true for the emails distributed by site 8.

While I did examine the privacy policy of every site used in this study, some of them offered, as so eloquently put by Professor Feigenbaum, “absolutely nothing.”
citefeigenbaum However, nowhere was warning given that an email address used on this site would be bombarded with ads for pornography, various weight loss and sexual drugs, less-than-credible mortgage applications, and in general “seedy” marketing.

Given the importance of volunteer privacy policies established in 2.3.3, it is disappointing to see so many vulgar and unsolicited emails. These types of lists represent the absolute absence of any control whatsoever; once an email address enters these lists, it will most likely continue to circulate until incoming mail starts bouncing. Even this is not a deterrent, since many times the goal of these emails is merely a website visit; the return-addresses supplied may not exist, or belong to innocent bystanders.

However, it is very reassuring to know that sites 5, 9, 1, 14, 4 and 11 held themselves and the other companies they shared any “consumer” data with to a strict code. In addition, sites such as 7, 13, 15, while providing a barrage of legitimate yet unwanted offers also seem to be committed to preventing the escape of customer data into the bottom rung of email offers.

A confounding variable here is the verifiability of confidentiality of trusted partners. The sites in middle quartile have relatively low incidences of unmatched headers, they do seem to share customer data with other firms. In some cases, it is likely that these other firms resold that data in violation of their agreements with the collecting websites.

For websites which allowed public view of the email address, email-harvesters may have come across the page and grabbed the un-mangled email. Projects such as Wpoison[11] are attempting to prevent this type of abuse. A single character hyperlink to a Wpoisoned page, surreptitiously hidden at the bottom of a page, would go unnoticed by the overwhelming majority of website users. An email harvester, however, would have its data collection irreparably polluted. Any sites that allow the public display of member email addresses have no reason not to implement such a system, given its low overhead.

On a much more positive note, after the initial round of data analysis, I contacted the webmaster and postmaster of all sites where $G1 \neq 0$ with the results of this experiment.¹⁰ With *one* exception, I received (relatively) prompt replies, all along the lines of “We are aware of the problem and are

¹⁰A considerable deal of this paper had been written before splitting the groups into their EQ and NEQ groups. The fact that for all sites where $G1 = 0$, $G1NEQ = G2NEQ = G3NEQ = 0$ shows that at least some organizations are acting in a responsible manner and holding their business partners to the same standards.

taking appropriate action.” Given the data Table 5 (or any of the tables, for that matter), it is readily apparent which site did not reply.

3.5 Further Testing and Analysis

While a number of Wpoison-style systems seem to be in use, readily available in perl, php or cgi, I did not come across any email seeding-tracking organizations. While confident of my data, at the time of this writing I have not checked the number of unexpected domain emails that arrived *after* the unsubscription grace period. Furthermore, due to the time constraints imposed on this project and other pressing matters¹¹, the observable window was fairly short. Additionally, the seed group of websites was rather small and I do not feel there were a sufficient number of comparable websites.¹²

Given sufficient future research, I now believe it to be possible to form a quantitative index of relative privacy levels that can be expected when a user signs up for a certain website. This is critical to the part of the privacy model that we will now flesh out.

4 Privacy Model Synthesis

The person with a new idea is a crank until the idea succeeds.

-Mark Twain

4.1 Overview

Relational databases allow for surprising cross-referencing ability. The concept of foreign keys and joins allow for massive data aggregation about specific individuals. For all data about a given individual “in the wild,” how hard is it to link all that data together. In the case of Ebenezer Q. Snodgrass III, probably not very hard. By measuring the dispersion of personal data and factoring in expected risk of a variety of actions, we can estimate how hard it will be to link that data together. Considering all data about a single individual to be in a relation, with the entire body of human knowledge as a gigantic database, we are actually measuring how hard it is to find candidate keys for an individual record in a combined relation of all knowledge. Obviously, only pieces of this key will be needed at any given time to extract useful data from information at hand, but by measuring the relative difficulty of collecting significant portions of the key, we can approximate how hard it will be to make use of any segment of it.

¹¹Pronounced “other classes.”

¹²e.g. slashdot.com and kuro5hin.com, m-w.com and dictionary.com, et cetera

4.2 Definition of Terms

4.2.1 Individual Bit-Descriptor

Suppose each individual is represented by the non-redundant bits $\alpha_0 \dots \alpha_n$. This bit-string contains such aggregate data as height, weight, ethnicity, parenthood, salient physical features, et cetera. Not all bits have equal significance, to allow for weighting of factors. Therefore, each bit or bit grouping has some constant θ_n by which it is multiplied. No two individuals will have identical bit-descriptors. A bit-descriptor is a unique id, many bits of which are known *only* to the individual in question. In addition, these bit-descriptors may change, albeit slowly, over time. Other entities may gain knowledge of all or parts of the bit-descriptor through observation, paper or electronic tracking, or via the direct consent of the individual. Put simply, the bit-descriptor can be thought of as a primary key in the grand relation of all human knowledge.

4.2.2 Trusted Entity

A trusted entity is an individual or other third party (e.g. a corporation) that has been granted consent to know/make use of a specific individual's bit-descriptor or some portion thereof.

4.2.3 Derived-Private Bits

In many cases, a trusted entity will be entrusted with many bit-descriptors or segments thereof, which may be vital to the way the entity operates. Under these conditions, the aggregate entrusted bit-descriptors form *derived-private* bits, which are appended to the bit-descriptor of the trusted entity. For example, customer mailing lists. While not initially part of the company's bit-descriptor, a customer database plays a vital part in the company's day-to-day operations, and thus the derived-private bits gain larger significance in the firm's bit-descriptor over time.

4.2.4 Privacy

Privacy is the expectation that no other individual shall collect, aggregate, decipher or otherwise have knowledge of an individual's bit-descriptor or any segment thereof without that individual's *explicit* consent. Furthermore, there is also the reasonable expectation that trusted entities shall not share trusted bits without the express consent of the issuing individual, and that trusted entities will only use those bits in the manner for which consent was granted.

In an extremely limited subset of cases, *knowledge* is sufficient, but in the general case we will prefer consent. For example, it would be unreasonable to suggest that all persons entering a baseball stadium sign a waiver; yet if that baseball game is televised, it is plausible during a "crowd view" that person's face will be publicly shown. This would fix that individual's position in both space and time (and hence disclosing some of their bit-descriptor) without their express consent.

Furthermore, there are sections of a bit-descriptor that are unreasonable to try to protect. For example, the bits that comprise a person’s outward physical appearance. Just by being in a public space, an individual clearly makes his or her “appearance” bits available to passers-by.

4.2.5 Security

Bit-Described Entity For the bit-described entity, in all likelihood an individual, the principal worries are that some of their bit-descriptor may be divulged without their consent, or that a trusted entity will either misuse its knowledge or that its knowledge will be compromised. In both cases, the individual no longer has knowledge of/control over which entities are aware of its bit-descriptor. Thus, security is the expectation that no bits shall escape an individual’s direct sphere of control.

Trusted Entity Due to the nature of derived-private bits, the worries for trusted entities are doubled. A trusted entity has both the responsibility of protecting the bit-descriptors it is allowed to use and the desire to protect its own bit-descriptor, which includes those derived private bits. Therefore, a trusted entity’s notion of security is the expectation that neither private nor derived-private bits (and by extension, the bit-descriptors they represent) shall escape that entity’s direct sphere of control.¹³

4.3 Incentives to Release Bits

Two individuals X and Y, with respective bit-descriptions $\alpha_0 \dots \alpha_n$. and $\beta_0 \dots \beta_n$, will differ by λ bits.¹⁴ We can then think of the differences between two individuals in terms of Hamming distance. The full bit-descriptors of individuals would, for obvious reasons, be fairly spaced out.¹⁵ Yet, we rarely (never) have the entire bit-descriptor to work with. A good example would be identical twins wearing the same outfit; to a casual observer, there is no discernible difference between the two. In this case, the observer has the exact same bit-descriptor segment for both of the twins, and $\lambda = \emptyset$.¹⁶

Therefore, bit-described individuals have an incentive to release sections of their bit-descriptors that will distinguish them from other individuals. If Ebenezer Snodgrass was a Yale student, the chances are good that his email address would be `ebenezer.snodgrass@yale.edu`. Yet, if a new student named Brad Kyle Rosen came to Yale, his email address would have to be `brad.k.rosen@yale.edu` to avoid

¹³In continuing the customer-database analogy, a disgruntled employee selling an entire customer-database to a direct competitor would be a tremendous blow to any firm: it would both disclose customer bit descriptors and make operations a bit more difficult, representing the disclosure of the firm’s derived-private bits.

¹⁴Or in the case of unequal length bit-descriptors, δ additions and λ bit flips.

¹⁵If you know someone *very* well, how likely are you to mistake them for someone else?

¹⁶Another anecdote involves a 9-year-old boy trying to take out a library book. The small local library was friendly to patrons who forgot their cards, and did lookups by first name, last name and town. Apparently a college aged “Brad Rosen” had taken out a book about sexually transmitted diseases, preventing his younger namesake from borrowing The Chronicles of Narnia. Had the library required another bit, such as home address or age, the situation would not have occurred.

confusion with mine. While “Kyle” and I would have $\lambda = 1$, this would be sufficient to differentiate between the two of us.

4.4 Lost Bits

While there are some sections of a bit-descriptor that an individual would not want freely available, such as his or her email address, there are other bits that no longer matter. For example, my telephone number last year was (203)-436-3036. While this is no longer current, this is part of my bit-descriptor. It really no longer matters if this is remembered or lost by any of the organizations I entrusted it to. However, if prosecutors were examining a database of prank 911 calls and they had not updated their database, a join on phone-number bits might implicate me. In this case, updated or lost segments are preferable to old ones. In the case of bank accounts, my current bank balance is part of my bit-descriptor, and of course, my bank is an entity entrusted with that knowledge, and until recently, both of my parents on my joint accounts. Had a malicious bank employee from 2.5 deleted these bits from the bank’s customer database, I would no doubt be in dire straits.

4.5 Bit-Segment Joins

Suppose that an individual X, whose name is represented by bit-segment S_0 has two email addresses, represented by bit-segments S_1 and S_2 . Furthermore, suppose that X enjoys gardening and is fond of playing devil’s advocate in political debates, represented by bit-segments S_3 and S_4 . Further suppose that X writes an email using S_1 to a public mailing list asking about nitrate fertilizer availability because he is sick of composting, represented by S_5 , comprised of the bits of S_1 and the bits of his message. Let’s further suppose that X writes another email, using S_2 to a political mailing list, in which he writes an exceedingly anarchistic diatribe in his role as devil’s advocate, represented by S_6 . Suddenly, X finds himself in serious trouble! Government domain databases would have access to S_5 and S_6 . Doing a join on S_0 on the mailing list rosters, they would soon find that X is an anarchist asking about nitrate fertilizer, a common ingredient in bombs. The government is unaware of S_3 and S_4 , either of which would have freed X from doubt. Instead, he is hauled off by the Department of Homeland security, questioned, embarrassed, and learns a good lesson. The bit-segments S_5 S_6 have come from the consumer domain (the companies running the mailing lists, presumably Yahoo!), yet they are being used by the government in a direct invasion of the private domain. In this case, the power of databases has allowed for an egregious violation of privacy.

4.6 Exposure of Bit-Segments

In the previous example, what if the government did not have access to the actual mailing lists, but instead, data that had been entered into a Federal database, purchased from a data warehouse. In this case, the privacy index developed in 3 can be used to estimate X’s risk, with the function:

$$f(x) = \delta_0 I_0 + \delta_1 I_1 \dots + \delta_n I_n \tag{1}$$

Where δ_i represents the expected risk (from our privacy index for site i) multiplied by the “danger” (I_i) of exposing the data submitted to site i . (I_i will be expressed as a function of the bit-segments entrusted to site i)

4.7 Bit-Segment Exposure Extension

While in the course of this paper we have only examined the databases of internet websites, traditional businesses also maintain relatively large customer databases as well. While it is trivial to extend the concept of the risk of bit-segment exposure to these areas as well, it is left to future research to determine appropriate metrics for examining this risk.

4.8 Other Database-Driven Bit-Risks

Using the MetroCard example cited in 2.3.1, there is the possibility of exposure of a huge number of bit-segments (fixed space-time position for dozens if not hundreds of datapoints). Activities such as throwing MetroCards in the trash at your residence would carry a higher risk than doing so in a public place. Burning or otherwise destroying the card after its usage would reduce the risk greatly, but not eliminate it. (Perhaps someone examined your wallet, or maybe you left your MetroCard on your desk and they wrote down the serial number and have a friend at the MTA)

There are many more such extensions to this model. As more data is available and they can be fleshed out, they will further refine the model to better reflect current privacy levels.

4.9 Summary

This privacy model relies on the ubiquitous nature of databases to approximate expected privacy. Extensions to this model to account for identity theft are possible, yet are beyond the scope of my current research. Database joins are crucial for the linking of information stored using a wide variety of bit-segments. Any increases in computing power increases the speed of these joins, and therefore the risk factor in publicizing certain bit-segments. The model adequately covers the three domains with its representation of derived-private bits, and the flow of information between the three domains as bit-segment exchanges. Threats to privacy in the status quo can be adequately explained by multiple joins on various bit-segments, yielding an aggregation of data about the individual best represented by those bit-segments. The inherent danger in the status quo that is easily demonstrated by this system is when data can vaguely incriminate or target the wrong individual, or merely annoy them by flooding them with unwanted marketing as their email, income and age bits are rapidly propagated into internet mailing lists.

5 Conclusion

If you put tomfoolery into a computer, nothing comes out of it but tomfoolery. But this tomfoolery, having passed through a very expensive machine, is somehow enabled and no-one dares criticize it.

-Pierre Gallois

In the status quo, databases and the information contained within them are being used in manners that our own government has agreed to be unacceptable. There is the constant risk that any private data released to another entity may become highly public, and there is currently no quantitative measure of measuring that risk. The bit-descriptor privacy model is a flexible model that can account for everyday phenomena and still provide an approximation for events that we do not expect to happen frequently. While this model is not meant to be a complete, it is the beginning of a robust model which can be extended and further refined. This tool will help retain some control over the runaway invasions of privacy in the *Database Age*.

References

- [1] Hacker hits up to 8m credit cards, February 2002. money.cnn.com/2003/02/18/technology/creditcards/.
- [2] Over one billion spam emails now blocked in one day by aol; members are helping to fine-tune anti-spam filters by reporting up to 5.5 million spam emails to aol each day, March 2003. news.cnet.com/investor/news/newsitem/0-9900-1028-20911774-0.html.
- [3] Ted Bridis. Justice department lifts fbi database limits, March 2003. www.siliconvalley.com/mld/siliconvalley/5472389.htm.
- [4] Joan Feigenbaum, March 2003. Professor of Computer Science, Yale University. Interview.
- [5] Simpson Garfinkel. *Database Nation*. O'Reilly, Cambridge, 2001.
- [6] John Leyden. On the microsoft ftp server leak, November 2002. www.theregister.co.uk/content/55/28252.html.
- [7] Mihn Luong, March 2003. Assistant Director International Security Studies, Yale University. Interview.
- [8] Jeremy Neuringer, March 2003. netNumina Consultant. Interview.
- [9] Oecd guidelines on the protection and privacy and transborder flow of personal data, 2002. www1.oecd.org/publications/e-book/9302011E.PDF.
- [10] Robert Sullivan. Domain registrar exposes customers, December 2002. www.msnbc.com/news/849290.asp.
- [11] Wpoison. www.monkeys.com/wpoison/.