

# Engineering Privacy in an Age of Information Abundance

Betsy Masiello and Alma Whitten

Google, Inc.  
1600 Amphitheatre Way, Mountain View CA 94043  
betsym@google.com alma@google.com

## Abstract

We now live in an age of information abundance. A near-infinite supply of information presents unique opportunities to tackle the most difficult contemporary social problems. The fields of medicine, energy and the environment, finance, governance, and even agriculture have already been transformed by information technology and are about to hit a second inflection point. The opportunities for improving the world are infinite, but a particular challenge is posed for individual privacy: much of the relevant information is inherently about people's behavior in society. In this paper we offer a conceptualization of privacy within the context of information abundance and present a set of engineering challenges that must be overcome to achieve it.

## Solving Contemporary Problems

It is quite possible that we are on the cusp of achieving a reality engineers thirty years ago fantasized about. Since the 1970's people have been talking about the Information Age, but everything that has come before pales in comparison to the information abundance we are beginning to experience.

Information today can be processed more cheaply and at a greater scale than ever before. The social Web is enabling crowd-sourcing of information for the smallest local governments, national governments, and the global community of Internet users. Mobile technology is advancing at a fast clip, enabling location-aware services and augmented reality. All of these trends will change the way we understand human behavior and society.

The beauty of this transformation is that it will create opportunities to solve the hardest problems facing society today. Contemporary problems are fundamentally different from those of the past. Yesterday we solved problems in relative isolation from each other; today's problems exist within infinitely more complex systems and are often interdependent. Solving them will require drawing intelligence from large amounts of loosely related information. As Hal Varian, Chief Economist at Google, has said, "the sexy job in the next 10 years will be

statisticians." The skills required to make sense of so much information are relatively scarce, but the demands for large-scale statistical analysis have never been greater, nor carried such moral weight.

## Monitoring public health

Among the most significant fields of research that will be driven forward by information abundance in the coming decade is public health. The breakthroughs will occur with the help of medical professionals, but they will not be possible without the application of statistical methods to vast quantities of information.

Identifying disease outbreaks at local levels is a micro-puzzle that, if solved, would go a long way in preventing pandemics. Imagine an ER doctor who at the start of his shift sees a patient presenting with flu symptoms. With the isolated information of this patient's chart, the doctor is likely to prescribe fluids and rest, even though more serious diseases might present with the same symptoms. What if he were able to see a visualization of the symptoms patients had presented with in the past 24 hours and the diagnoses? If one or two had presented with flu symptoms and been diagnosed with meningitis, he might treat this patient proactively to limit further spread of the disease.

Similar information collection and analysis across the health value chain could also be done. Simply monitoring the number of ER admissions in a region or town could reveal a lot about the status of local disease outbreaks. Or consider the intelligence a pharmacy could draw by monitoring purchasing patterns for over-the-counter medications or prescriptions. Imagine if we were able to monitor school and work absenteeism across a region, not just at a particular school or company, in real-time.

Applying information analytics at scale across the value chain would prove tremendously useful in monitoring disease outbreaks as they are happening. But what if we could predict the conditions that heighten risk of disease spread before they occur? If it's been a particularly rainy summer in a malaria-infested region for example, we can

expect a higher rate of malaria infections. Do similar but less obvious weather patterns exist for other diseases? Or are there other conditions that meaningfully impact disease transmission and spread?

Studying aggregate health records may make even less obvious improvements. Researchers in South Africa examined the health records of HIV and AIDS patients in an attempt to understand factors that influenced the progression of the disease. They discovered that there was a strong link between vitamin B consumption and delayed progression to AIDS. How many more HIV patients could we treat if we were able to do so with a cheap, commodity supplement like vitamin B in comparison to expensive antiretrovirals? This is just one example, but where there is one there are many. Examining in aggregate complete individual health records of specific populations might reveal information that would lower the cost of treatment, or identify the source of a disease, or uncover a cure for currently incurable conditions.

It's clear that making progress against the most pressing public health challenges we face will require collecting and sharing information that is fundamentally about the health of individuals. Like most information analysis of this scale, the aggregate information is more important than the individual, but disambiguating individuals from each other is critical. From a privacy point of view, it's clear this present risks, but few would argue that we should halt progress on these problems.

### **Saving the environment**

Climate change is increasingly a well-accepted phenomenon that may have dramatic consequences for the planet. Managing it requires, among other things, getting a better understanding of electricity consumption over time and geography. There are two aspects to this: the energy suppliers' ability to predict energy demands to optimize renewable sources, and the consumers' ability to make tradeoffs about his energy use that incorporate knowledge of the system-wide energy use. The Smart Grid is poised to go a long way toward solving this problem.

Most of us have no idea how much energy we use in our homes on a daily basis, which devices and appliances consume the most energy, or which times of the day put the highest or lowest overall demand on the grid. Studies have shown that when people have access to these types of information, they typically reduce energy consumption by 5-15%. In isolation, the impact is small, but in the context of the complex system that is national energy demand it matters a lot. If half of American households cut their energy demand by 8%, it would be the equivalent of taking 10 million cars off the road.

Imagine how much more efficient we can make our energy consumption as a whole if we collect and share information with each other. Is it possible that this approach could apply to other scarce resources as well, for example water or landfill space? Is a communal approach to aggregate information about resource consumption an early answer to the tragedy of the commons? With respect

to any common good, it seems feasible that more information about the complex system in which it is being used and restored would help all of us manage it better.

### **Governing**

There has been tremendous interest in increased government transparency. There are at least two reasons, the first of which is that transparency about the behavior of government officials will more easily uncover corrupt practices. The second, and possibly more powerful, is that by making government information more widely available there is an increased likelihood that previously unforeseen intelligence will be drawn from it.

In the US and UK efforts to make government data more broadly accessible and shareable have led to development of third-party tools that simplify tracking government actions, or even notifications of small concerns like potholes in local roads. We can also imagine sites that enable local citizens to overcome collective action challenges and improve their local community through volunteerism.

The last US Presidential election was, as we have all heard about, run with significant use of social networking sites that enable voters to build communities around issues they care about. None of this would be possible if not for the ability to collect, share, and analyze information.

### **The importance of allowing for unexpected sources of information**

It seems quite possible that some of the most innovative approaches to tackling these problems will be done with information that was collected to solve a different problem. Google's Flu Trends and Domestic Trends are two examples of approaches to some of the problems we've described that have been built using aggregate search query data, not originally collected with these innovative applications in mind. It's unlikely that any user three or four years ago entered a Google search thinking, "this data point will be one piece of predicting flu outbreaks." It's equally unlikely that many engineers had thought of this possible application when building the search engine.

It is impossible to predict the ways in which a set of information might be analyzed to understand a loosely related problem until that information exists. The wonder of the technical trends leading us toward information abundance is that they lower the cost of information collection and processing such that having an exact outcome in mind a priori isn't always necessary.

Thinking about privacy in this context, one natural reaction might be concern. The collection of so much information raises the risk of unauthorized access to information and subsequent privacy violations. But quickly revisiting the moral weight of the social problems that we are facing, we should appreciate the opportunities created by analyzing aggregate information. In the context of this new world, we face an uncomfortable question: is

minimizing information collection always a practical approach to privacy in the 21<sup>st</sup> century?

## Conceptualizing Contemporary Privacy

If we agree that the nature of these problems is hitting an inflection point because of information abundance, we also will agree that restricting collection and analysis of information would slow progress toward solving them. For those of us concerned about individual privacy, the resulting challenges are daunting. Information collection has historically presented a privacy risk because once recorded, information can too easily be intentionally or accidentally reapportioned in ways that violated the privacy of the subject.

Privacy is not dead, though. As has happened time and time again in the past century, our expectations of privacy are just adapting to the new environment today's innovation is building. This constant evolution has always made it difficult to settle on a firm definition of privacy, and information abundance does not change this dynamic. However, a few concepts have remained with us through the years that are a useful starting point in conceptualizing the issue.

**Reputation.** Many of the approaches to thinking about privacy have at their core the idea that a person should have at least a modicum of control over their reputation. This has been fashioned as having an ability to regulate information about oneself, control over one's personal information, and the ability to limit access to one's personal affairs. It is implicated by concepts not traditionally associated with privacy, such as gossip and lies, both of which invoke control, or lack thereof, over the accuracy of source information. Paradoxically, extreme transparency is increasingly a method individuals are choosing to exercise privacy as control of one's reputation.

**Ephemerality.** The digital revolution changed memory forever. The impermanence of source information has long been a de facto protector of privacy. Much like the game of telephone, information morphs as it is verbally retold. Papers get lost, degrade over time and can be destroyed with some ease. But information on the Internet is made reproducible and searchable instantly, preserving the original source indefinitely. Caches such as the one maintained by the Internet Archive keep information publicly accessible even after the original site is taken down. On the Internet as we know it today, impermanence is often not a practical expectation, yet human nature yearns for the ability to express thoughts momentarily, without the threat of permanent access.

**Secrecy.** Private doesn't necessarily imply individual, it can also imply the privacy experienced within a group who will keep information shared therein secret. Trust is critical to maintaining secrecy, without it the threat of gossip or inadvertent disclosure is ever-present. Similarly, security often goes hand-in-hand with secrecy. If access to the group is not secured, privacy cannot be guaranteed. Secrets

may be kept only to ourselves, in which case their privacy is dependent on security of their storage place. Anonymity can convey secrecy by removing the identifiability of the information.

**Contextual integrity.** Privacy has until recently been approached as a somewhat binary issue: information is public, or it is private. Recent scholarship has identified the fallacies in this approach, indentifying contextual integrity as a necessary condition for maintaining privacy. What follows is the recognition that privacy exists on a spectrum. Information we're comfortable making public in one context we may wish to keep private in another.

We will almost certainly need to adapt our expectations to the practical implications of each of these conceptualizations if we want to capture the opportunities created by information abundance. If information is made ephemeral, it becomes less useful in understanding complex problems over time. If kept secret, it is useful only to a limited group of people in understanding complex problems, and there's no reason to believe that group will have incentives to investigate these problems. The proliferation of information about us, particularly when secrecy and ephemerality are harder to achieve, leads us to feel like we're losing control over our reputations. Digital information persists across contexts, and few mechanisms exist to signal one's preference for contextual integrity.

Another way of framing the problem is to investigate the concern that is at the core of these motivations. Our hypothesis is that it is a concern about how others are going to treat us once information about us is revealed. "Will I be socially isolated or mocked?" "Will I be convicted of a crime I didn't commit?" "Will this affect my employment eligibility?" "Will the health insurance company deny my claims?" "Will the nature of my relationship with this person change in an irreversible fashion?" At the core of these concerns is the desire to feel in control of our own fate.

These concerns actually have very little to do with technology, but paradigm shifts in how we communicate and share information have happened at a faster clip than social norms or laws can adjust. The solution is not, however, to suggest we all just "get over it" and adjust more quickly. Instead, we should be focusing our engineering efforts on finding solutions for this broad spectrum of privacy concerns.

## Enabling Meaningful Control

There are several engineering challenges ahead that make this an exciting time for privacy research. Among the most pressing demands is to give users control over their personal information, particularly when information abundance is the norm. Today, we tend to think about this control in the context of social networking, but in this new world it applies to far more information than that you share with other users. Once individuals have meaningful control over their information, they can choose to be left out of

aggregate information analysis while still enjoying the value of information services. Without control, users can often choose not to use an information service or to allow information about their interactions to be analyzed, but rarely can they do both.

It is very difficult to object to the principle that users should control their personal information, but technical solutions are not necessarily obvious or easy to achieve. We can easily overlook tradeoffs when talking at a high level about the ideal outcome, but limitations still exist that prevent deploying solutions immediately.

**Providing access to unauthenticated data.** Anonymity exists in context. A shopper may be anonymous to the cashier, but if the cashier were asked at a later point to re-identify that shopper in a different context and with more information, he might be able to. The privacy community has recognized the risks presented by re-identification methods applied to digital information, but re-identification is not the same as authentication. Anonymous information will always carry some risk of re-identification, if combined with enough additional data sources.

This is a critical point worth extra emphasis. For the majority of positive uses of aggregate information analysis, including those outlined earlier in this paper, a certain amount of uncertainty is entirely acceptable and will not degrade the utility of the analysis. Many of the most pressing privacy risks, however, exist only if there is certainty in re-identification, that is if the information can be authenticated. As uncertainty is introduced into the re-identification equation, we cannot know that the information truly corresponds to a particular individual; it becomes more anonymous as larger amounts of uncertainty are introduced.

Unauthenticated data brings this concern front and center. We would all like to understand what might be learned about us through re-identification of unauthenticated data we've left behind as we surfed the Web, but practically speaking giving a user access to unauthenticated data presents more privacy risks than it does provide comfort. If there is any uncertainty that information is linked to you, for example you are using a shared computer and executing unauthenticated searches, presentation of that unauthenticated search history may reveal private search queries executed by one of the people with whom you share the machine. How, then, do we meaningfully support access to and control of unauthenticated data that though anonymous in context is theoretically re-identifiable if given enough additional information?

A related issue arises out of the desire to increase contextual anonymity by breaking linkages between sequential actions. As discussed, the benefits of information abundance can only be achieved by analyzing information that is fundamentally about humans. In other words, sequential linkages are often important to preserve, but the identity of the person associated with these actions is rarely necessary to preserve. Can we design a way to

maintain linkages between sequential actions while decreasing the likelihood of re-identification?

**Preventing abuse.** Among the most critical reasons to collect and analyze information about use of a system is to detect and stop abuse of that system. In the case of a search engine, this practically speaking means having the ability to provide quality results that are not spam, to guarantee to advertisers that they will not be victims of fraud, and to prevent denial-of-service attacks. Much of the information collected to achieve these aims is anonymous, in context, but understandably its collection raises privacy concerns.

An important point about detecting abuse is often lost in privacy debates that rely on opt-in and opt-out rhetoric: some types of information collection are optional, while others are, at least today, not optional. Privacy sensitive users would like systems to exist that don't rely on information collection so that they can trust their exchange with the system is ephemeral and secret. How can we achieve this ephemerality and secrecy without limiting our ability to detect and prevent abuse?

**Managing the paradox of choice.** One way to give individuals more control over their reputation is to provide fine-grained choices about information collection and use. In theory, this is an admirable goal, but in practice overwhelming the user with too many choices can shutter the usefulness of a system. Sometimes we want every option available to us, and other times we want to get directly to the utility of the system. These preferences vary by person, context, and even particular instances of interaction.

This is fundamentally a usability challenge. How can we maintain the utility of the systems we've all grown to love while increasing the control we have over how we interact with them?

**Conveying the benefits of aggregate information analysis.** Most if not all of the information collection and analysis that will solve society's most vexing problems depends only on anonymous, aggregate information analysis. The benefits will accrue to all of us, but there is a free rider problem that, if overwhelming, could negate the utility of information. Enabling control should be the top priority, but the close second should be conveying to the individual what benefits will accrue as a result of their participation in the information analysis.

Traditional medical research provides an interesting, but admittedly imperfect, analog. Information about our health is considered among the most private, yet when we fall ill with an incurable disease we rarely hesitate to participate in research studies that depend on collecting and analyzing information about our health. Trust is an important part of this, but so is the ability to convey the utility of participating in the research. We feel personally affected by a disease, and want to be a part of saving others from the same fate.

We are all affected by public health crises, environmental crises, financial crises, and corrupt government. When we participate in information collection writ large, we aren't aware of what solutions we might be

contributing. Starting with even the most simple of examples – that of high quality search results depending in part on information about what other users have found useful – how can we convey in a visceral way to the individual user that collection of their search query will contribute to higher quality information for everyone?

**Supporting social signaling.** Maintaining contextual integrity online often depends to a large extent on the actions of our peers. Posting a photograph on one social networking site does not necessarily imply we want it posted on a popular blog, or even on the public Web at all. We depend on others recognizing our preferences to maintain the contextual integrity that is important to us. And yet, the tools available for doing so today are inadequate.

Think back to lunchtime in elementary school. Imagine a bully takes a private note of yours and passes it on to one of the popular students in your class. He has the information at his fingertips and the ability to break its contextual integrity by reading it aloud to the cafeteria. Imagine he sees you looking fearful, your cheeks turning red; perhaps you're even on the verge of tears. We hope that such a situation would end by the student returning the note to you unread, having picked up on the social signals you were sending that its contextual integrity had been compromised.

The inability to convey these social signals online has been pointed to as one cause of online bullying, but it is also at the core of enabling privacy on the social Web. The solutions pursued so far on the social Web have generally been to technically enforce privacy preferences by giving users the ability to limit access to information. This method doesn't support a social contract though; to do that, we should begin investigating ways to effectively signal privacy preferences to others.

**Enabling self-representation.** It is desirable to give individuals as much control as possible over their self-representation online. We all deserve the ability to correct misinformation that is spread about us; in the offline world we want opportunities to counter unfair gossip and there's no reason to think it should be any different online. The ability to change digital information, which in some cases might amount to censorship, must be balanced against preserving free speech.

A common discomfort people experience is seeing information about them appear in search engine results. One reaction people have is the desire to remove information from the index, and in some cases there is very good reason for wanting to do so. It is a burdensome process, though, as it requires contacting all the source URLs and requesting the information at the source be changed, first. Another reaction people have is to overwhelm the search index with information they themselves create, a sort of radical transparency. But what about the situations in which neither method works?

The case of "dog poop girl" in South Korea illustrates the type of problem that can arise. The information that resulted in her reputation being marred was factual; it

would have been a restriction on free speech to censor its dissemination. But what can the young woman do to repair her reputation? What tools are available to help her do so? It is difficult to start fresh with the global reach of the Internet; we ought to find ways to enable people to start fresh when circumstances call for it.

This is a collective action problem in that no single information provider can develop a solution on their own. There is also a cultural element involved in balancing free speech against self-representation in these situations. And finally, there can be a legal question of defamation involved. Of all the challenges outlined, this is among the most challenging because reasonable people may disagree about the desired outcome. For that reason alone it is a challenge that deserves our attention, thoughtful discourse, and problem-solving skills.

## Conclusion

We began this paper by exploring a few of the most pressing contemporary social problems we face and the ways in which information abundance could hold the key to solving them. Few privacy discussions start from this context, which we believe does the issue a great disservice. As it's typically discussed, privacy and information collection can often seem at odds: to protect privacy, put restrictions on the type of information collected, the ways in which it is collected, and the length of time it is retained. Addressing the means and rationale behind data collection is an important first step, but the information abundance context highlights where the transformational solutions for privacy will be found.

In a world where information collection is the norm, what we need are not just a set of principles or policies, but also a set of technical solutions that give users meaningful control. Few of these solutions are at our fingertips today. We don't have a mechanism to absolutely protect against any and all re-identification risks associated with unauthenticated information. We don't have reliable ways to prevent abuse without collecting some baseline information about interactions with a system. Our current solutions to provide users choice are often overwhelming, or so simple as to limit real choice. It is not obvious how to convey the potential, often unknown, value of aggregate information collection to an individual user. Technology lacks the ability to convey the social signals we are accustomed to in our everyday lives. As a global society, we face a difficult and complex tradeoff between an individual's control over self-representation and another individual's ability to express themselves freely.

None of these challenges have simple solutions. They cannot be overcome simply through application of principles, policies and procedures – those alone only get us half of the way there. The path forward must ultimately be blazed by the engineers building tomorrow's information services and by the statisticians making sense of the information. Until we address some of the challenges described here, we will struggle to move past

the tension we struggle with today between, on the one hand, limiting information collection as a means of protecting privacy and, on the other hand, making advances against our most pressing social problems through data analytics.

## References

All for Good. <http://www.allforgood.org>

Borelli, Adam. August 12, 2009. "Announcing 14 Geo Challenge Grant Recipients." *The Official google.org blog*. <http://blog.google.org/>

Ginsberg, Jeremy and Matthew Mohebbi, Rajan Patel, Lynnette Brammer, Mark Smolinski, and Larry Brilliant. November 19, 2008. "Detecting influenza epidemics using search engine query data." *Nature*.

Granada, Hannah Choi, Jon Creyts, Anton Derkach, Phillip Farese, Scott Nyquist, and Ken Ostrowski. July, 2009. "Unlocking Energy Efficiency in the US Economy." McKinsey & Company, Global Energy and Materials.

Google.org program pages. <http://www.google.org>.

JMP Blog. September 17, 2009. "Live from Malcolm Gladwell's Speech at Discovery, Innovator's Summit." Retrieved from <http://blogs.sas.com/jmp/index.php?archives/243-Live-from-Malcolm-Gladwells-Speech-at-Discovery,-Innovators-Summit.html>.

Manyika, James. January, 2009. "Hal Varian on how the Web challenges managers." *McKinsey Quarterly*. [http://www.mckinseyquarterly.com/Hal\\_Varian\\_on\\_how\\_the\\_Web\\_challenges\\_managers\\_2286](http://www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286)

MySociety projects page. <http://mysociety.org>.

Nissenbaum, Helen. "Privacy as Contextual Integrity." *Washington Law Review*, Vol.79, No. 1, 2004.

Rijsberman, Frank. October 26, 2009. "Will genomics help prevent the next pandemic?" *The Official google.org blog*. <http://blog.google.org/2009/10/will-genomics-help-prevent-next.html>

Solove, Daniel. 2007. *The Future of Reputation: Gossip, Rumor and Privacy on the Internet*. Yale University Press.

Youth for Human Rights International. "Human Right #12: The Right to Privacy." [http://www.youthforhumanrights.org/watchads/view/psa12\\_h.html](http://www.youthforhumanrights.org/watchads/view/psa12_h.html)