

An Exploration of Distributed Social Networking

Monica S. Lam
Computer Science Department
Stanford University

May 13, 2009

Abstract

Based on current trends, this paper predicts that a monopoly will emerge that owns our personal data in the cloud, and discusses ramifications of loss of privacy and fair-market competition in the software industry. We argue that a technically superior alternative would give ownership of the data back to the individuals based on a distributed, egalitarian design with open interfaces that allow a choice of storage vendors and software vendors. We present an architecture and a prototype that sports some of these features. We urge systems researchers to rise to the challenge to create an economically viable, user-friendly, egalitarian data system. Topping the research agenda is the monetization of privacy, and how companies can make money out of personal data without compromising privacy. Other research topics include creating self-managed personal servers, high-level programming systems to enable easy development of distributed social-networking applications, and techniques to safeguard data confidentiality.

1 Introduction

More and more of our personal data are making it onto the web every day. From applications as pedestrian as word processing to social networking tools such as Loopt, which allow one to share their GPS location with friends, the web is supplanting the classic Personal Computing paradigm. The smart phone is accelerating this trend. Users expect to be able to view data produced on the desktop while on the go. Platforms such as Google's Android synchronize personal information with a central server to allow easy access from both the phone and the PC.

Centralized application services have many positive properties. They are easy to use. What could be simpler than a single user interface to both use and configure your personal data? They make it easy to share. If you and your friends use the same site, it is trivial to share photos, messages and possibly a SuperPoke or two. It is fun and easy to discover new friends, to discover who your friends have as friends or to reconnect with old friends. They are always available anywhere on any device from the smart phone, to the desktop, to the work computer; you can keep up with your friends anywhere. They keep getting better all the time. Users do not need to worry about software upgrades as the application provider automatically updates the software as needed. Third party application developers have, through independent development, made these platforms more useful and fun than ever before.

All of this freedom does come at a cost however. The risks created by centralized service providers is worthy of concern. Lets fast-forward to the year 2020 when you find yourself in quite a quandary:

MyFace Corporation (NASDAQ:FACE) is a publicly traded company that runs the social network that everybody on earth is using. MyFace's stock price is in the dumpster at a 52-week low. A savvy group of investors crunch the numbers and realize that MyFace is worth more as a direct mail marketing service than it is as a social networking application.

After reading this news, you immediately log into your account and press the "Export Profile" button followed by the "Cancel Account" button. Well, you would have, except these buttons are nowhere to be found. You go to bed slightly uneasy, but presume that the new group of investors will run the company with the same level of integrity as the previous team.

The next day you wake up to the sound of your phone ringing with an offer regarding your “recent interest in Guitar Hero XLV.” Moreover, your mailbox is overflowing with letters, your inbox is 3GB over capacity, and you have 392 new text messages on your iPhone 7G.

In the following storm of litigation, the company is rendered financially insolvent. A bill to bail-out and partially nationalize MyFace is debated by congress but, it is already too late. MyFace goes off line and your data disappears forever. Yeah, you wish you had control of your data.

This story is but one of many outcomes possible if current trends continue. Throughout the rest of this paper, we will analyze the risks created by the centralization of personal data, attempt follow current trends to their logical conclusion and propose one possible alternate architecture, PrPI, that allows users to keep control of both their public and private data. We believe that research into systems such as PrPI is fundamentally necessary in order to, at the very least, elicit a response from the current application service providers.

2 Monopoly!

2.1 An Odd Sense of Foreboding

Today, while cloud services allow users to be more productive, there is trouble on the way. As more and more of our information leaves devices we control, it will become impossible to maintain the privacy and ownership that storing data locally guaranteed.

From the point of view of one attempting to covertly access data, which are ostensibly private, it is hard to beat a centralized repository that is taken care of by a third party. In this model, participants freely volunteer even the most private of information to this third party. Many types of spying are possible. Personalized marketing is used to finance many of the current service providers. Marketers depend on the results generated by data-mining to properly target information to individuals and not to waste users’ time with inappropriate and irrelevant advertisements. This can be far more effective, but invasive, when able to use personal secrets. How alarming would it be to get a prescription sample in the mail when data-mining program ascertains you may have a disease before you even start to feel sick? This could be discovered by your flight history, passenger log and the seating arrangement on the plane. How about a telephone call from an estate lawyer when data-mining discovers that, given your genetic profile, your health may soon start to deteriorate? How much worse would it be if this were shared without discretion?

2.2 Bad Moon Rising

Looking forward, it seems probable that surrendering personal data to third parties will, with time, lead to a monopoly. In current social networking and data storage applications, people must create multiple logins and enter the same friends and data multiple times. This is clearly inefficient from a user’s standpoint. Another problem is that users will want to be on the same platform as their friends, family, coworkers and acquaintances. This network effect is an enormous force in producing a dominant player. The gravitation generated by others creates a lock-in; you may not want to keep using the service, but in order to keep in touch with friends you must stay on the site. Further binding you to the application service provider is the fact that it is incredibly difficult and time consuming to pull your data off the site. It is probable that users will become inured to the lack of control and associated loss of privacy.

As a dominant player emerges, it will be able to offer more functionality. By having more personal data and more users on a single platform, new classes of applications will become possible. Data that was once scattered about the web will now be accessible in a uniform method. This will feed back into the success of the dominant player. Facebook is now offering much more than just a homepage with instant messaging, photo sharing and user-developed applications. Social networking applications will be even more powerful when merged with traditional data such as email. Due to high infrastructure costs, many companies will be building on this platform.

Companies that try to develop their own infrastructure will run into scaling issues if they succeed. These scaling issues are driven by the network-effect whereby users bring their friends on board. Finding it unable to keep up with an exponential increase in usage, they may need to be sold to companies that have both

the financial and infrastructure resources necessary. Formerly independent providers such as YouTube and Flickr have already been assimilated into larger companies. Following these trends to their logical conclusion, a monopoly will soon emerge.

2.3 Rage and Ruin

By definition, a monopoly that owns our data also owns the commanding heights of the personal computing universe. The company will not only hold the data but also define how they are accessed. They will completely supplant the traditional personal computing paradigm by becoming the personal computing platform.

Examples of “evil” from companies that exert 100% control of a platform are legion. One only needs to remember Netscape’s long descent to obscurity when Microsoft introduced Internet Explorer to understand the forces at work. Apple’s iPhone is another good example. Competing with any iPhone software product from which Apple derives revenue is strictly verboten [?]. Additionally, Apple disallows third party usage of certain APIs which gives them an advantage in developing immersive applications (although developers seem to be given some leeway) [?]. In the context of personal data, the company may choose certain partners to whom special data access will be given. These partners may be given faster access to data, on-site compute resources and may even get a hand in designing the next generation of interfaces. This closely mirrors the current net-neutrality debate: should carriers be able to give priority to some data while possibly slow-boating other data?

When choice disappears from a market, so does any incentive for a company to compete. This nadir can lead to all sorts of problems. A company in this position could hold our personal data hostage. Perhaps a monthly service charge is in order for a company that used to offer its services free of charge. Advertising could become more invasive as, absent market forces, the company no longer feels the need to keep data private. Through poor management, technological atrophy or government regulations meant to assuage the first two concerns, the company could fail to make a profit, fail and take our data with it. There is also a slight possibility that the monopoly would be so slow at reacting to a new challenger that could itself reach market dominance.

The ultimate irony will be that a company that made our lives better by offering free services now rules the world. It is quite possible that, in 2014, 30 years late, a version of Orwell’s prediction may finally emerge. Rather than the government, which is already viewed with suspicion, it is instead a for-profit company that knows all.

3 Is Resistance Futile?

Given the huge economic incentive for companies to become the dominant owner of our personal data, it is a formidable task to provide an alternative to an omniscient, personal data monopoly. Not only is it necessary to provide the same advantages as a centralized service, any alternative would have to make economic sense in order to attract sufficient capital to fuel its development. In the following, we first describe an approach that is technically superior to a centralized service architecture. Then we present some research ideas that can make the alternative economically viable.

3.1 Distributed Personal Cloud Infrastructure

As the monopoly has not yet been realized, the same weaknesses that are driving services to consolidate may also leave room for a new approach. Today, most of our personal data are still stored on our PC, instead of having different kinds of data stored on different web services. On average, a Facebook user only has 66 photos online, a tiny fraction of what we own and continually acquire. The obvious alternative to centralized services is distributed systems where users retain ownership of data; they can decide between choose to store their data on devices they physically possess, or store them in the cloud if they so please.

Just like distributed PC computing has overtaken mainframe as a more economic solution, a distributed social networking infrastructure is fundamentally more efficient than centralized services. A distributed system can take better advantage of the locality of references and economy of scale, and allow individuals the autonomy to upgrade their equipment and capacity independently. Even today, an average household

in America has more than sufficient processing power, storage, and network bandwidth to enable sharing of all of their own personal data with all their friends and relatives. In a centralized model, however, the scaling cost to cover the American population is immense in terms of storage and computational capacity, network bandwidth, cooling, real estate, and administrative costs. There is no such thing as a free lunch. The consumers in the end are paying for the inefficiency indirectly, because centralized service providers need to pay for the infrastructure while seeking profits.

Typically, technology incubated by Computer Science professionals in universities and companies eventually make its way to consumers. Distributed systems have not made this leap. Consumers have the same need to share media with friends over the Internet. We envision that personal servers of tomorrow may become as prevalent as today's personal computers. A passive, encrypted copy of the data in the server can be backed up into the cloud—note that ISP's could cheaply provide such a service since it is not actively computed upon. Thus, personal servers may not be as distant and impossible as it may seem. Tivo's are media servers that are found in many consumer households; game consoles are participating in distributed computations such as Folding@home. Obviously we still have a long way to go before today's social applications - performing functions such as sharing GPS locations, media, or discovering friends - can be correctly translated to personal server architectures. The key is to create open high-level distributed programming interfaces and frameworks that enable independent software vendors to create distributed applications that run across these servers.

3.2 Monetization of Privacy

The existing cloud services are very lucrative. In order for any alternative to succeed, its revenue generation potential must be greater. Privacy, the key factor in our new design, must be monetizable. Any platform we design must take advantage of the privacy to (1) make possible a new class of viral applications and (2) preserve and even enhance the ability of advertisers to make a profit.

Without privacy, an entire class of financial and medical applications will not be accepted. Imagine a medical advice program that processes a family's medical records to recommend medical checkups, diets, exercise regimens, and life insurance purchase strategies. The alternative would be for each family member to maintain an up-to-date and unencrypted copy of their data at a service provider, which is cumbersome and undesirable for privacy reasons. In fact, privacy is also useful for applications involving interpersonal relationships, a particularly viral category. While it is generally accepted that the younger generation has less qualms over making personal information public, few would be willing to make public their negative feelings about other individuals. A blind date application is a good exemplar of the socially-aware category of applications due to the fact that it requires privacy between members of a potential couple, but more information must be released to the mutual friend who is playing matchmaker. The most personal of information is necessary to make this work in any reasonable way and, this is the type of information that a person may feel uncomfortable releasing to an omniscient third party to be placed, indelibly, in a centralized database.

To compete successfully against the marketing dollars fueling centralized services growth, a new alternative must provide even better targeted marketing opportunities. Given that data stored on personal servers are assumed to be private, we may be comfortable with it storing the history of our entire digital life, namely photos, documents, calendar, contact information, communication, purchase and browsing histories. While it is a marketer's dream to get access to a person's entire digital profile, the challenge lies in utilizing one's data in targeted advertisement and marketing research scenarios without giving away privacy. Here are two scenarios that illustrate this possibility. Imagine walking into Macy's with your cell phone and the list of sales items are presented to you in order of your likely interest based on your aggregate digital history. Privacy is maintained, for example, by having Macy's present to your cell phone the list of all sales items. The phone, utilizing private data from your personal server, performs list ordering, not just at Macy's but across all physical and digital stores, even accounting for upcoming birthdays of friends as stored in your personal calendar. To allow market research, consumers can opt-in to participate in market research where, in a sampling basis, only a small fraction of one's personal information is sent to marketers after anonymization. People could be incentivized monetarily to participate in such a program.

4 PrPl Personal Cloud Infrastructure file:///figures/arch1.jpg width=3.5in

In this section, we describe the design for PrPl (PRivate-PuBLic), an egalitarian system infrastructure to address privacy and monopoly issues outlined in the previous section. Fig. 1 shows an overview of PrPl. Each user has a Personal Cloud Butler, a server that is owned and controlled by the user. The butler maintains a centralized index of all data belonging to a user, including different data types like pictures, video, audio, email, relationships and GPS locations. The butler provides an aggregate view of all the data, though actual data could be physically stored on different computing devices like laptops, desktops, online services or mobile devices. The Personal-Cloud Butler is so-named to suggest that it is entrusted with the most confidential information, and it carries out services with discreetness, often with very high-level direction. The PrPl architecture has several benefits, as outlined in the following sections.

A unified view over a federated storage system. We embrace the consumers' desire to take advantage of the many available web services. To that end, the PrPl infrastructure creates a federated storage system out of existing services and storage devices. The user can continue to take advantage of free e-mail services and access data he has uploaded on Facebook. At the same time, he can choose to place sensitive data only on his home PC, with possibly an encrypted backup in the cloud. Once the user registers various data repositories, the infrastructure automatically tracks changes, requiring no more human intervention. Our infrastructure presents a unified, location-agnostic view of the data in a user's personal cloud. The butler may choose to cache or replicate data across storage devices.

Uniform and fine-grained access control. Instead of uploading data to be shared to specific sites, a user would simply specify, at as fine a granularity as desired, with whom he wishes to share his content. Once a user is authenticated, an application running on his behalf can access all of his data as well as the data shared with him by his friends. This makes it easy to develop applications that require access to large collections of information owned by different people.

High-level semantic queries. The PrPl butler creates a semantic index of all the user's data and puts it into an RDF store, using ideas from the semantic web [?]. This allows the user to tag data once, and reuse it in many different applications. A user's butlers can synchronize with another user's butler to view or query information to which it has been granted access. Many of the emerging social networking applications, especially those on mobile devices, can be framed as a semantic query on data owned by a group of friends. Many of the top 50 Android applications have that flavor [?].

The PrPl butler needs to be highly secure because it indexes all of a user's private data. We therefore prefer it to be placed under the user's direct control and not allow any third party direct access to data held by it. For example, users could potentially run the butler service on a highly reliable home gateway. An encrypted copy of the data may be stored by the butler in the computing cloud (e.g. Amazon S3) for backup purposes.

Implementation We have implemented a prototype of PrPl to test out these ideas, and have successfully written some social applications using it. In our prototype, the federated storage system can incorporate data from IMAP email clients, public information from Facebook, Flickr, Yelp, and Google using their web service APIs, file systems on any PC or laptop, and rich media on the iPhone, Android phone, etc. We have written applications as semantic queries on top of this infrastructure, and find that several interesting social applications can be expressed succinctly, typically in a few hundred lines of code.

5 Research Challenges

In addition to the need for privacy monetization approaches discussed above, there are many technical challenges that must be overcome before distributed cloud computing can become a reality.

How to make it easy to create distributed applications by mere mortals?

For distributed computing to take off, we have to make it dramatically easier to create distributed applications. The overall goal of creating a personal cloud infrastructure is to elevate the level of programming. Map-Reduce is one example how hiding the complexity of parallelism has enabled rapid development of thousands of parallel applications by non-experts [?].

As we build distributed infrastructure, the challenge is to create open API's that allow easy programmatic access to data, allowing interoperability with existing systems. Open, decentralized frameworks like OpenID [?] and OAuth [?] in the areas of decentralized authentication and authorization respectively, serve as useful subsystems for infrastructure development. Further, emergence of standard ontologies like those

from Project Nepomuk [?] will allow consistent organization of data across disparate systems. Central to the development of this infrastructure are decisions on how the data are to be represented, stored, and cached as an optimization.

We are exploring the development of distributed Datalog as a high-level language for writing distributed queries for social applications. Datalog is a declarative logic programming language; distributed Datalog would further hide from the programmer the complexity of determining where the computation is to be performed and the logistics of coordinating a distributed computation. High-level constructs need to be created to take advantage of domain-specific characteristics. For example, it may be necessary to cap queries, based on time or the quality of the results, on large amounts of data owned by friends of friends.

How to provide security and prevent leakage of private information?

Personal servers storing financial and medical records spanning our entire digital life are likely to be magnets of attacks. Furthermore, enabling distributed social network applications while also guaranteeing privacy is a challenging problem. It is critical that we allow users to specify fine-grained access control without undue difficulty, for example, by using metadata, tags or semantic attributes. In addition, techniques such as *information flow control*, possibly implemented at the language, compiler, as well as the operating system [?] would be useful safeguards against coding errors.

6 Concluding Remarks

Online services are acquiring personal data at an astounding rate, as users find it appealing to have their data available anytime, anywhere, and on any device. We discussed how this may even spur the development of a monopoly.

We proposed a distributed platform that retains the core functionalities of a centralized service with the additional advantage of returning ownership of the data to the user. The existence of a distributed solution offers consumer choice and puts pressure on centralized services to treat our data with the care and discretion we desire. A distributed, technically more efficient, platform may even supplant centralized services if privacy can be monetized to generate new classes of viral applications and more effective targeted marketing.