# Supporting Collaboration in the Era of Internet-Scale Data

Cameron Marlow
Facebook, Inc.

## 1 Introduction

Reproducibility is an often-cited and valid concern of the research being performed by many corporate research programs, such as the one I work with at Facebook. We consistently run experiments on millions of active users using proprietary systems, gather results on data infrastructure at massive scale, and produce reports which distill this process into a few lines of context. It is no wonder that many papers from internet research labs are returned with comments to how the results are interesting but entirely irreproducible.

This effect is just one symptom of the growing gap between the instruments available to researchers studying social computing, human-computer interaction, recommender systems, and auction theory, among others. On one side of this divide are academics, depending on shared data sets and infrastructure to enable the collective advancement of science, cooperating with Institutional Review Boards and beholden to funding agencies. On the other side are industrial researchers, utilizing proprietary data and infrastructure to driving science forward, maintaining privacy and Terms of Service (TOS), and beholden the goals of the corporation. It is hard to say how wide this gap is, but clear that the computational power of the likes of Google and Facebook continue to grow.

When asked to produce a position paper on challenges in studying technology-mediated participation, I thought naturally to address the questions I most often get from academic researchers: can I have some data? Can I crawl the users on my university network? Perhaps run a query on your databases? At the same time, papers are regularly published which violate Facebook's TOS, expose users privacy, and without any regulation by ethics review boards. In this paper I hope to describe what Facebook has to offer, expose some of the challenges we currently face in engaging with academia, and propose some possible solutions which allow for direct collaboration while upholding all legal and ethical guidelines.

## 2 Anatomy of an Internet-Scale Social Research Tool

Before addressing the outstanding challenges to sharing research agendas with academia, I will first introduce some of the basic components of a modern internet-scale research

apparatus. This is not meant to be a prescriptive list, but rather a list of tools that an industrial researcher can take for granted.

**Instrumentation.** From human-computer interaction to search relevance, the value of large-scale internet research starts with data quality and robustness. While the implementation varies between systems, most social media sites capture rich data about their users. First and foremost, users describe themselves in *dimensional data*, including demographics, geography, and other personal features. Second, the bread and butter of social networking services is *structural data*, in the form of personal relationships, interests, ratings and groups. Third, participation in various activities is captured through *action logs*, connecting users to rich behavioral data such as search queries or written reviews. Finally, a number of less rich, such as page views and navigational clicks may be summarized as *transactional data*. With adequate instrumentation and an active product, the possibilities for research are greatly expanded.

**Data Infrastructure.** Another critical component in large-scale internet research is the computational infrastructure in support of offline data analysis. While early internet companies depended on relational databases for storage and analysis of log data, more companies today are opting for grid-computing systems that enable scalable computation built on top of commodity hardware (such as Hadoop). What this means for researchers is the ability to construct rich queries that cover terabytes of data in minutes, in an environment that also supports standard business intelligence, analytics and engineering needs of the rest of the company.

**Experimentation.** The final component of an internet research platform is a rich and robust framework for performing experiments on active users. In order to fully test hypotheses, users can be randomly assigned to different conditions, exposed to different experiences, and compared across a number of dimensions. Experimentation can be performed anywhere from back-end ranking algorithms to treatments of the visual user interface, all utilizing the same random sample of users. These controlled trials can lead to the generation of a fundamental understanding of all aspects of user behavior.

## 3 Challenges

Given the value of engagement between industrial and academic researchers, there are a number of challenges to supporting collaboration across the firewall. Social media applications usher in a new era of rich, dynamic social interactions with precise privacy controls that allow for sharing to limited audiences. The explosion of semi-private data on the web leads to a number of challenges around policy regulation and explicit data sharing. These

issues are the center of debates on ethical and fair research practices across many academic communities [4].

**Sharing data.** The most straightforward method for collaboration is through the exchange of data sets, typically anonymized by obscuring the identity of the user. Depending on the nature of the data, it can be challenging to produce privacy-preserving transformations that do not obscure the value. In the case of user action logs, such as web queries, the content of actions can easily give away the identity of a user [3]. More surprisingly, unique patterns in the structure of social networks can be used to unlock the identities of the individuals involved [2], making the most interesting and valuable data of social media the most difficult to share.

**External research.** When a user publishes information to a social media service, they have expectations about who will see this information and how it will be used. The Terms of Service is the most common way to communicate to users and third parties how that information will be used, and who will have access to it. At the same time, the ease of access to individuals in these services has led to an explosion of research using the data available to individual researchers on Facebook. Projects that would clearly be rejected by an IRB or TOS review are being performed at an increasing rate, and being published in highly reputed academic publications. Policing these policies is nearly impossible, but academic publishers should begin holding researchers accountable in the same manner as Institutional Review Boards. By upholding Terms of Service, academic can collect data and produce research that supports user privacy. At the same time, for those researchers trying to carefully observe the conditions of a TOS, even the most simple data under user consent will be considered a violation, making external research nearly impossible.

## 4 The Future of Collaboration

Most of the challenges listed in the previous section arise from the inability to share data without consent, transfer of data at scale or the misinterpretation of legality of various data collection activities. The easiest form of collaboration involves the direct engagement between universities and corporations, under contract. This approach has been very successful, and will no doubt continue to be the most useful form of interaction. The downside of this approach is the overhead of interaction and resources needed to support an individual outside researcher. In order to scale corporate data and facilities to a larger audience, the following approaches will be valuable.

**Shared infrastructure.** Ignoring the data provided by internet companies, the easiest way to support academic research is to level the computational playing field by providing shared grid computing for use in approved projects. The NSF, IBM and Google recently

entered into a partnership to provide such a service, allowing NSF-funded researchers to utilize thousands of processing units for purely academic research. At the same time, companies such as Amazon have been extending reasonably priced grid computing and even pre-installed public data sets [1], which are within the budget of many academic projects. The downside of this relationship is simply that it provides infrastructure, but with a Bring Your Own Data policy that ignores the problems that inspired the construction of the system.

**Streamlined consent.** A common theme in social science researchers is the marrying of survey responses to user behavioral data. For instance, an online survey may ask a user to provide estimates of the strength of their relationships to friends on a social network, and these data may be compared to communication patterns within the service. This type of research requires both the consent of the user under standard ethical practice, as well as the rights from a service to store data. Coupled with a solicitation system, this type of practice would enable researchers to easily deploy surveys, collect data, and report without the involvement of the service. Standard bias calculations for survey nonresponse could enable wide ranges of research while supporting user privacy and upholding data retention policies.

**Aggregative queries.** The most involved solution to collaboration within the cloud would be to enable researchers to perform queries on corporate data, using corporate infrastructure, but using privacy-preserving data-mining techniques. For instance, given a language for specifying queries on a large-scale data set, facilities could be provided to guarantee anonymity and diversity among the output of such a query. Similarly, aggregations could be performed, preserving the identity of individuals but allowing for distributions of features across the population. While this method is difficult, and also still prevents many types of data access (such as large-scale whole networks), it presents an engaging approach to collaboration, albeit unlikely to exist in the near future.

## 5 Conclusion

As industrial settings produce more and more powerful data and computational tools, it becomes increasingly important to enable collaboration between academic and corporate researchers. The introduction of rich, dynamic social data represents a goldmine for social scientists, but collaborating in environment that promotes user privacy is an increasingly challenging task. In this paper I outlined some of the challenges facing the liaison between internet-scale corporations and academic researchers, as well as some potential solutions.

## References


[1] Amazon, Inc. Public Data Sets on Amazon Web Services (AWS). URL `http://aws.amazon.com/publicdatasets/`.

[2] Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 181–190, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. doi: http://doi.acm.org/10.1145/1242572.1242598.

[3] M. Barbaro and T. Zeller. A face is exposed for AOL searcher no. 4417749. *New York Times*, 9:2008, 2006.

[4] Nathan Bos, Karrie Karahalios, Marcela Musgrove-Chávez, Erika Shehan Poole, John Charles Thomas, and Sarita Yardi. Research ethics in the facebook era: privacy, anonymity, and oversight. In *CHI EA '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pages 2767–2770, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-247-4. doi: http://doi.acm.org/10.1145/1520340.1520402.