Research Statement

Lars Backstrom
Cornell University
lars@cs.cornell.edu

January 24, 2009

My primary research interests are in data mining and machine learning using large-scale datasets, with an emphasis on Web information, social computing applications, and on-line social networks. These datasets are becoming more and more numerous, and as the Web's reach continues to grow, it is important to understand these datasets for two reasons. First, better understanding of the dynamics of the systems generating the data allows us to improve the systems. For instance, if we better understand how a social network grows, a social network provider can better serve its customers when they try to expand their friends lists. Second, an in-depth understanding of the data allows us to leverage it for a variety of purposes. For instance, by looking at where search queries come we can discover the reach of various topics, ideas, and opinions; and by understanding how new ideas spread on a social network, we can improve marketing efficiency. I am particularly interested in developing new methods and algorithms to deal with these large datasets, answering the subtle and nuanced questions that require a huge amount of data and novel methodology to deal with.

Social and information networks One area of particular interest to me is the analysis of large-scale information networks. In particular, I have focused on datasets stemming from online social networks. Most early analysis of social networks fell into one of two broad categories. Sociologists typically studied these networks in great depth, but were limited to small-scale networks because the methods they used involved questionnaires and interviews and hence did not scale well to thousands of people. On the other hand, in computer science the networks being analyzed were typically much larger, but the questions being asked were consequently much simpler. The focus was on simple network properties like degree distribution and diameter – interesting properties, but somewhat limited. In *Group formation in large social networks: membership, growth, and evolution*, we tried to bridge this gap by using techniques from computer science, and in particular data mining and machine learning, to ask more complex questions and tease apart subtle distinctions, all on large-scale datasets which have only become available very recently.

In particular, we looked in depth at questions related to groups or communities of individuals from two datasets. Over time, these groups grew and evolved, and our goal was to tie the network properties to this evolution. Using techniques from machine learning, we were able to discover which features of the social network were most important in predicting whether a person would join a group or not. We found, for instance, that the most important factor is how many friends one has already in a group, and we were able to quantify the nature of the dependence of joining probability on number of friends (it is roughly logarithmic). At a finer level of detail, we found that if one's friends are well connected, that also increases one's chance of joining: a person with three friends who are all friends with each other is more likely to join than a person with three independent friends. In addition to looking at individuals, we also looked at groups as a whole and examined how the network structure of the group was related to its growth rate. Here we found that to accurately predict group growth, one must take into account a broad range of network features and that, while no one network feature was an accurate predictor on its own, when put together we were able to build an accurate model for predicting the growth of communities.

In this work we were able to answer some questions which have been discussed and theorized about for many years. One particular contribution of this work is that it showed the exact shape of influence curves. With so much data, we were able to determine with great confidence the probability that a person would join a group as a function of how many friends he or she had in the group. While influence has been discussed at length in the past, it was not possible to evaluate it with such confidence until recently because the datasets and methods were not available to do so.

More recently, in *Microscopic Evolution of Social Networks*, we looked in great detail at how individuals behave in these social networks. What factors influence when an individual will form a new link, and to whom will that link go? We found, for instance, that almost all new links go to friends of friends, and that by using some network structures and recency effects, we can determine which of all the friends of friends' a person is most likely to link to next.

By continuing to apply techniques from computer science to these large social networks, I think there are many questions that we can answer which were impossible until only recently. As more and more data becomes available, we can continue to devise a more and more general and in-depth understanding of these systems. With this understanding, we can both build better, more usable systems, and also put the systems we build to better use.

Spatial and geographic embeddings of information In addition to social networks, I am interested in developing a better understanding of other information systems. A good example of this is search engine queries. While an individual query contains little or no useful information, the entire corpus of billions of queries contains a wealth of information. For instance, one can easily tell when the moon was full by looking for high volume periods of the query 'full moon'. A recent example of this which caught worldwide attention is *Google Flu Trends*, where query volume, combined with geolocation, allows Google to report which states have the highest flu rates.

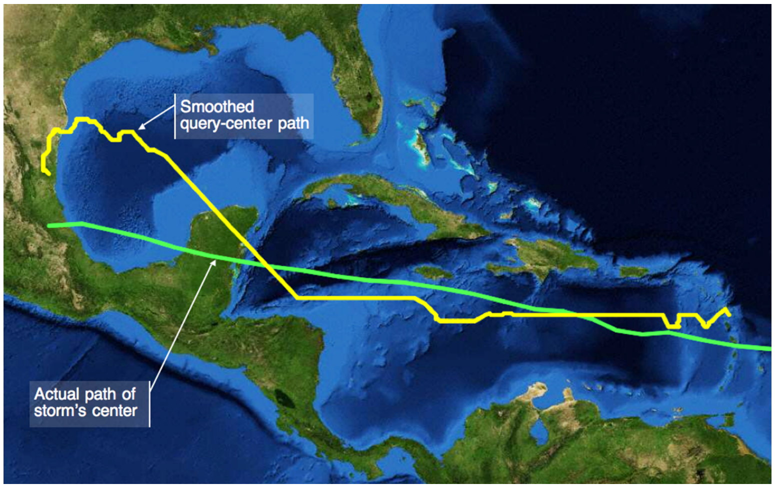

Figure 1: The path taken by the hurricane, compared to the path of the query center (note that the center is sometimes in the sea because of the queries surrounding it).

In my own work, I have looked at what information can be gleaned from queries combined with

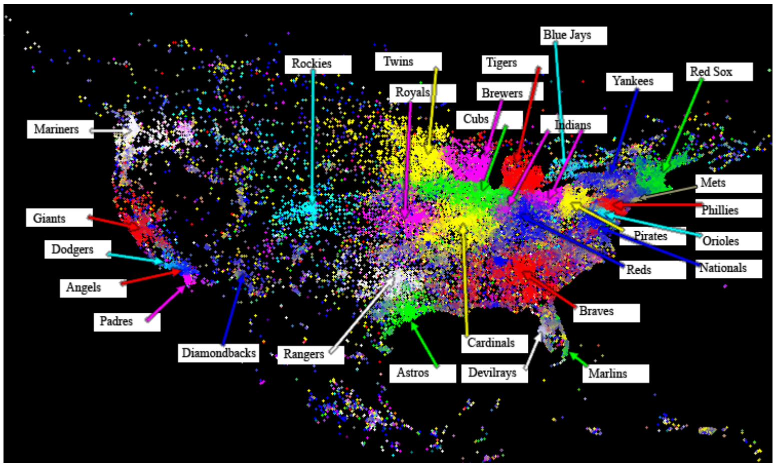

Figure 2: The influence of many baseball teams falls along state lines.

their location of origin. In *Spatial Variation in Search Engine Queries*, we found that, despite the inaccuracy of finding the origin of a query, we were able to learn a great deal from the queries and their approximate locations. At the simplest level, we were able to automatically discover the locations of most landmarks and large (but local) organizations. For instance, using this dataset with the algorithms we developed, we could correctly identify the cities of origin for all major newspapers, sports teams, and many national parks. Furthermore, we discovered a number of geographic relationships which are less obvious, such as which social networks are popular in which parts of the world (Facebook is particularly popular in Ontario, for instance). By looking in more depth, we are able to track queries through both time and space, learning how their geographic profile changes over time. A nice example of this comes from the query Hurricane Dean, where we were able to recover an approximation of the hurricane's path by looking only at where the related queries originated.

Not only can we learn about individual items, but by comparing different queries we can compare influence between various competing products. As an example, the accompanying figure shows queries for baseball team names, coloring a map of the US according to which team(s) were dominant as queries in that area. This work was done with colleagues from Yahoo on full Yahoo query logs, and it was featured in discussions with Yahoo senior management (as well as in the search industry trade press) for its potential as a way to improve the targeting of search advertising.

Currently, I am exploring geographic directions in the context of photo collections. Here, there is an opportunity to combine the photos' locations with textual information (i.e. tags) and image content to better organize photos. With collaborators at Cornell, I currently have a paper in submission along these lines.

Privacy in on-line data Finally, while there is much to be learned from all of these large-scale datasets, one must also be careful how they are treated. Much of the most interesting data contains sensitive personal information, which must be handled delicately. Individuals have typically entrusted this information to some caretaker, and in order to continue to be able to study this data in aggregate, we must be careful to respect that trust. In many cases, it is not obvious how this sensitive information might be leaked out. While it seems rather obvious in retrospect that releasing search queries, even with some information removed is a recipe for disaster, it is harder to see what might go wrong in other circumstances. In particular, it is not at all obvious what might go wrong if a company were to release a large, but completely unlabeled, social network. With millions of unlabeled nodes it is, at first glance, difficult to imagine how any single individual could be identified in the network.

In *Wherefore Art Thou R3579X?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography*, we looked at exactly this question. We imagined that a company released a large, unlabeled network, where links represented some relationship between two individuals (for instance, had two people ever exchanged emails). We showed that, even with this seemingly innocuous data, an attacker could, quite easily, compromise the privacy of some individuals in ways that they would likely find unacceptable. We gave simple, efficient algorithms whereby an attacker could create just a few strategically placed nodes and links before the release of the 'anonymized' data and then discover the identity of other individuals in the network. By discovering the identities of two specific individuals of interest, an attacker could then determine if those two individuals had a link between them, a clear violation of privacy.

Because I think data analyses of social content is important, I would like to focus on both sides of the privacy issue. I think it is important to point out what could go wrong when we are not careful, but at the same point, I would like to work on systems of data release with privacy guarantees under some reasonable models of privacy.

In the future, I would like to focus on improving the functionality of these systems through the development and use of algorithms which extract useful information from the human generated data. By understanding the systems in greater depth and by extracting useful information from the generated datasets, I believe that there are many opportunities to improve social networks, search engines, news sites, collaboratively generated content, and more.