SOCIAL DATA PROCESSING

By Devavrat Shah, Professor of Electrical Engineering and Computer Science; Member, Laboratory for Information and Decision Sciences (LIDS), Institute for Data, System and Society (IDSS); Director, Statistics and Data Science Center (SDSC)

Background. What went wrong with the election polls in the 2016 U.S. presidential election? How can the online activity of the population help curate better life experiences for all? Can we utilize online personas for reaching out to individuals in a targeted manner? What about predicting the demand for espadrilles this summer? Or ranking the performance of your favorite sports team? And what happened to the promise of using collective wisdom for stopping the spread of ''fake news'' on Facebook?

Answers to all those questions depend on our ability to process "social" data to extract meaningful information. For the past few decades, and even more so recently, everything online is being recorded. If aliens came to earth and inspected the social data — generated by us as a society (not by machines) — they would learn, for instance, that Patriots' Day is when the Boston Marathon is held.

Put another way: Social data presents us with an enormous opportunity for making data-driven decisions for better living, more efficient operations, more effective policy making, and overall uplifting of societies. Here, data is the enabler. Access to social data has been democratized; the key to success lies in the ability to process it so that we can extract meaningful information from it. This makes it feasible for someone like me as an "ivory-tower" academic to carefully think through a solution, test it out, and then have a chance of making an impact in the real world — and, in the process, advance the foundations of statistics and machine learning. In that sense, social-data processing presents an unusual, potentially once-in-a-generational, opportunity that can lead to a remarkable convergence of academia and industry.

Challenge. The standard approach for data-driven decisions following statistical decision theory is to use an appropriate model that connects data to decision variables, helps make desired predictions, and eventually facilitates optimization over decision choices. Here, data is generated by humans, so modeling social behavioral aspects is essential. Any social scientist can attest that modeling human behavior is an extremely intricate task and the resulting models can be highly context-dependent. That makes it especially challenging to come up with effective, meaningful models. The only hope is for gaining access to *lots* of social data to decipher the right model for a particular interest from a large class of models. In other words, we need a model that is flexible enough to capture a wide array of social scenarios. But the model must be sufficiently tractable so that with *enough* data it can capture the ground truth faithfully, and it is important that such a system can computationally scale along with the data.

In short, the key intellectual challenge is in finding a sufficiently flexible model for social data that is both *statistically* and *computationally* tractable. This is a major challenge, and its successful resolution can have substantial impact on all the previously mentioned scenarios — and many others.

Turning Weakness to Strength. To progress toward such a grand challenge, it is essential to identify the properties of social data that are ubiquitous across a variety of scenarios and that can be captured to develop meaningful models.

We have identified one such property: social data is (or should be) *anonymous*. That is, from the data-processing perspective, it should not matter who has generated the data. To put it another way, the overall conclusion should remain invariant if we re-name the individuals who have generated the data. For example, the results of a democratic election should not change even if the voters' names change, as long as the total number of votes for each candidate remains the same. In the same way, the popularity of a specific style of espadrilles does not depend on which specific individuals bought them, only on how many pairs are being purchased.

Anonymity seems like a constraint or a weakness from any angle you look at it.

After all, anonymity and privacy protections restrict the type of information that we can mine from data. But we derive strength from this apparent weakness. It will help us address the challenge of developing tractable and flexible models for social data.

Mathematically, anonymity can be viewed as the underlying "probabilistic model" having a certain "exchangeability" property. A remarkable development in mathematical statistics, starting with the work of de Finetti in the 1930s with further developments in the 1970s and 1980s, provides a crisp non-parametric characterization for such models: the Latent Variable Model. We utilize the Latent Variable Model for social-data processing for a variety of scenarios, including some of those discussed in the questions that opened this article.

Taking the First Steps. We start by examining the question of designing personalization or recommendation systems such as those used by Netflix, YouTube, Amazon, and Spotify. Here, the goal is using the history of an entire population's preferences to predict which movies, music, books, or other products that individual consumers may like and that they have not already experienced. On one hand, the question is: What is the best algorithm to design for that end goal using the non-parametric model emerging from exchangeability? On the other hand, that question has been with us since the dawn of the e-commerce era.

There is a popular algorithm, called Collaborative Filtering[0], that has been with us from the start and that continues to be used due to its simplicity and empirical success. In a nutshell,

the algorithm embodies the following time-tested insight: If your friend likes a new movie, and that friend's tastes are similar to your own, you will likely enjoy the new movie as well. Such a simple, intuitive — or, may I say, *social* — algorithm has been used in practice successfully but with little understanding of why it works. Ultimately, the goal is to understand the Collaborative Filtering algorithm and, in the process, try to find ways to improve it — and, if feasible, achieve the best performance.

In our work[1], we have precisely addressed this question. We use the non-parametric model arising from exchangeability to study this question. We find that the Collaborative Filtering algorithm, somewhat miraculously, is solving the right statistical problem. In a nutshell, it implicitly performs "local" approximation for the non-parametric functional model underlying the data *without knowing the function!* The nature of the algorithm leads to accurate learning when the available data is sufficient enough. However, this is not the "best" performance in terms of the data required for accurate learning.

The above framework provides systematic means to improve upon the Collaborative Filtering algorithm. To begin with, in the regime where a lot of data is available, a simple improvement to the Collaborative Filtering algorithm can lead to superior performance (see Figure 1). This improved algorithm has the best-known performance for the most generic model class as argued in[1].

A well-known limitation of the Collaborative Filtering algorithm in practice is its inability to work well in the presence of very sparse data. Imagine the scenario of YouTube where 300 hours of video are uploaded every minute, or a shoe retailer where completely new designs of espadrilles are introduced every season. There is very little data about new shows or espadrilles across the population. In the context of Collaborative Filtering, you may find that none of your friends has watched the new shows or tried the new espadrilles. Therefore, the algorithm may not able to provide meaningful recommendations.

In a recent work[2], we extend the Collaborative Filtering algorithm to overcome this sparse data limitation by using the following "social insight" guided by the non-parametric statistical model: *your friend's friend can be your friend; or more generally, if you and your friend have similar preferences, and your friend's friend has similar preferences to your friend, then your friend's friend may have preferences similar to yours.* The resulting "iterative" Collaborative Filtering algorithm turns out to have (near) optimal statistical performance in sparse data regime. And it's remarkably simple.

Where to Go from Here. The non-parametric Latent Variable model is useful beyond the setting of recommendation or personalization. Over the past decade, as a community, we have developed solutions for a variety of scenarios. Notable ones include:

(a) finding aggregate ranking over a collection of choices such as teams, players, or faculty candidates by synthesizing data available in the form of partial rankings or preferences such as pair-wise comparisons[3];

(b) finding accurate answers to questions such as those collected through polls and surveys or crowd-sourcing platforms based on noisy answers[4]; and

(c) finding communities in a society based on noisy pair-wise interaction data[5].

It turns out that for each of these scenarios (and more), the Latent Variable Model is a less restrictive, or more flexible, model while being tractable. Subsequently, this provides a way to develop a data-processing algorithm for a wide variety of social settings simultaneously. For example, I am very excited about our ongoing project, where we use the Latent Variable Model for time-series data in high-dimension for accurate forecasting in real time.

The Latent Variable Model has applicability beyond social data (cf. see [1] and [2]). For example, it can help de-noise image data by viewing images as 3-order tensor of RGB values and connecting to a Latent Variable Model. An example of the efficacy of the Collaborative Filtering algorithm for de-noising image over academic benchmark images is described in Figure 2.

Summary. Social data presents us with a tremendous opportunity. To realize the opportunity, it is essential to develop statistical models "universal" enough to faithfully capture a broad class of "social" scenarios; and inference algorithms for such models that are statistically and computationally tractable. This is a grand challenge, because modeling social behavior that generates data is extremely hard. The non-parametric Latent Variable Model naturally arising due to the *anonymity* property of social data is a promising candidate in making progress towards this challenge, especially given the initial progress made.

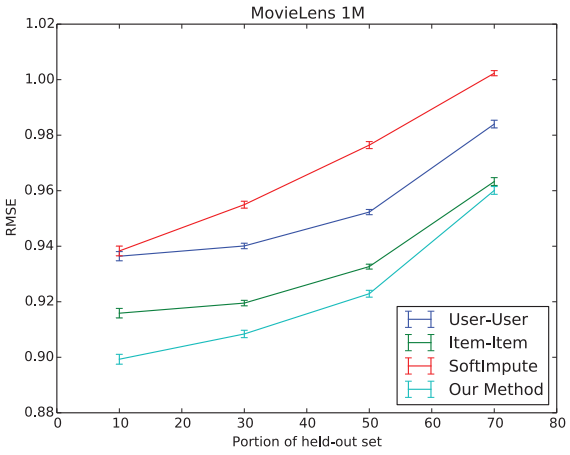

Figure 1. This is an experiment representing the performance of various recommendation algorithms using the MovieLens dataset. The performance of the algorithm is measured in Root-Mean-Squared-Error (RMSE) — the lower, the better. Algorithm performance is evaluated for different fractions of test data (the rest of the data is training). The orange curve corresponds to spectral method (soft-impute), the blue curve corresponds to user-user variant of Collaborative Filtering, the red curve corresponds to item-item variant of Collaborative Filtering, and the purple curve corresponds to the improved Collaborative Filtering algorithm using the Latent Variable Model.

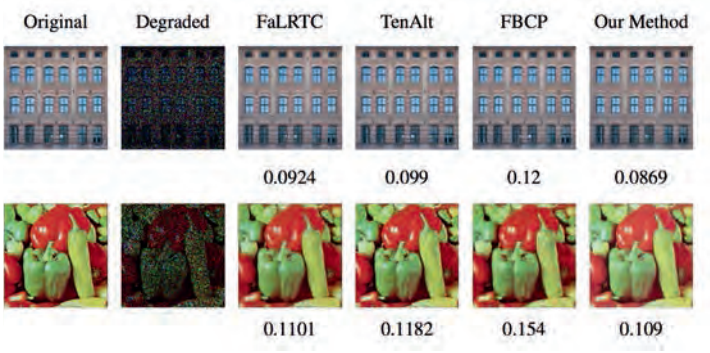

Figure 2. Recovery results for two images (building facade and peppers) with 70 percent of missing entries under different algorithms are presented. The last column corresponds to our algorithm, which is based on Collaborative Filtering. The performance is measured with respect to Relative Squared Error (RSE) — again, the lower, the better.

References

[0] D. Goldberg, D. Nichols, B.M. Oki, and D. Terry. Using Collaborative Filtering to Weave an Information Tapestry. Communications of ACM, 1992.

[1] C. Lee, Y. Li, D. Shah, and D. Song. Blind Regression: Nonparametric Regression for Latent Variable Models via Collaborative Filtering. Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS), 2016.

[2] C. Borgs, J. Chayes, C. Lee, and D. Shah. Recommendations for Sparse Datasets via Similarity Based Collaborative Filtering. Preprint, 2017.

[3] S. Negahban, S. Oh, and D. Shah. Rank Centrality: Ranking from Pair-wise Comparisons, Operations Research, 2016. Preliminary version in Proceedings of NIPS, 2012.

[4] D. Karger, S. Oh, and D. Shah. Budget-Optimal Task Allocation for Reliable Crowd-Sourcing, Operations Research, 2014. Preliminary version in Proceedings of NIPS, 2011.

[5] C. Moore. Computer Science and Physics of Community Detection: Landscapes, Phase Transitions and Hardness, Bulletin of EATCS, 2017.

"Social data presents us with an enormous opportunity for making data-driven decisions for better living, more efficient operations, more effective policy making, and overall uplifting of societies."

NEXT-GENERATION NANOSYSTEMS: A Q & A WITH MAX SHULAKER

By Anne Stuart | EECS

Max Shulaker, an expert on nanosystems exploiting emerging nanotechnologies, joined the EECS faculty in the fall of 2016. He is the Emmanuel E. Landsman (1958) Career Development Assistant Professor of Electrical Engineering and Computer Science and a principal investigator for both the Microsystems Technology Laboratories (MTL) and Research Laboratory of Electronics (RLE). At MIT, he is starting the Novel Electronic Systems (NOVELS) research group.

He received bachelor's, master's, and PhD degrees in electrical engineering from Stanford University. His PhD research on carbon nanotube-based transistors and circuits resulted in several firsts:

- the first digital systems built entirely using carbon nanotube field-effect transistors, or FETs (including the first carbon nanotube microprocessor),
- the first monolithic three-dimensional integrated circuits combining arbitrary vertical stacking of logic and memory, and
- the highest performance and highly-scaled carbon nanotube transistors to date.

At MIT, Shulaker is launching an experimental research program aimed at realizing his vision for the next generation of electronic systems based on transformational nanosystems, leveraging the unique properties of emerging nanotechnologies and nanodevices to create new systems and architectures with enhanced functionality and improved performance.

Shulaker was interviewed in his Building 39 office, which overlooks the ongoing construction for MIT.nano, the new nanoscale fabrication and characterization facility scheduled to open in 2018. He spoke about his past and current research, his new experimental program, and his early experience at MIT.

Q: How did you become interested in nanosystems and nanotechnologies?

A: I got interested in this area in a class on digital systems that I took freshman or sophomore year at Stanford. The professor talked about trying to make a computer out of carbon nanotubes. It seemed like a crazy idea, but I asked the professor if I could help, and he said "yes." That started close to a decade of working on carbon nanotubes.

The more things I did in the lab, the more excited I became about the technologies. That's one reason I always encourage undergraduates to get involved in research — you never know when you will find your passion.

This work was exciting to me because it spanned all layers of the computing stack. I began focusing on the materials and carbon nanotube synthesis. Then we started looking at circuits. Then we started looking at systems. Then we starting looking at new applications. And now, my own PhD students are working on projects that span all those layers as well. They have to work on the new materials to build new circuits to enable new systems to demonstrate new applications.

When you add up all of the benefits across all these different layers, you aren't talking about 10 or 20 percent benefits anymore, but instead gains exceeding several orders of magnitude. This work has the potential to make a huge difference in the world, and it is why I — and my students — are so excited to be working on it.

"I began focusing on the materials and carbon nanotube synthesis. Then we started looking at circuits. Then we started looking at systems. Then we starting looking at new applications. And now, my own PhD students are working on projects that span all those layers as well."