THE WIDE WORLD

Researchers at Cornell's Social Dynamics Laboratory use online networks to study human behavior at a once-unimaginable scale

By Beth Saulnier

Let's say you wanted to study millions of people worldwide with the aim of understanding how mood fluctuates throughout the day—and from day to day, month to month, and season to season. How would you go about it?

If you had unlimited resources, maybe you could deploy a vast army of trained researchers to observe your subjects in shifts, 24/7—but that would be a huge logistical undertaking, and the privacy issues would inevitably get thorny. You could send out surveys and ask people to recall how they felt from minute to minute—but those kinds of retrospective recollections are notoriously unreliable. You could hand out devices that allow people to report their moods in real time—but what if that very action interfered with how they're feeling? And what if they just forgot to do it?

On a practical level, in other words, such a large-scale study was impossible—until the advent of social media. Those networks, which have upended so much of everyday life, have opened up new opportunities for researchers in the social sciences and other fields to study how people think, feel, behave, interact, form opinions, and more. Among the prime movers in this research revolution: Cornell's Social Dynamics Laboratory (SDL), whose headline-generating work has included a 2011 study leveraging Twitter data to explore the mood cycles of more than two million people in eighty-four countries ›

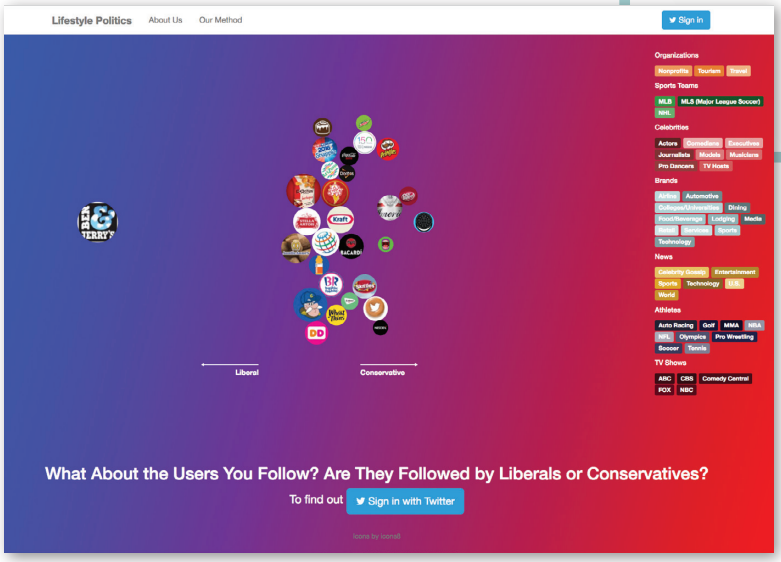
ILLUSTRATION: JOHN HERSEY/THE ISPOT

Latte Loving Liberals?

In 2004, a now-infamous ad by the conservative Club for Growth decried Howard Dean's supporters as a "tax-hiking, government-expanding, latte-drinking, sushi-eating, Volvo-driving, *New York Times*-reading, body-piercing, Hollywood-loving left wing freak show." The idea of the "latte liberal" stuck—and it eventually inspired researchers in the Social Dynamics Lab to study how well the red-blue divide correlates to seemingly nonpolitical choices like what hot beverage to drink. Tapping existing data from a large national social survey, grad student Daniel DellaPosta and colleagues reported in the *American Journal of Sociology* in 2015 that there were indeed marked contrasts between lifestyle preferences for liberals and conservatives. They built on that work by studying five million people who follow the Twitter feeds of 553 current and former members of Congress, and looking at which lifestyle feeds those people also follow; Ben & Jerry's and Starbucks are favored by liberals, for example, while Chick-Fil-A is a conservative darling. Among the team's more recent discoveries: beer is favored not only by conservative men, but by liberal women.

The idea of the "latte liberal" stuck—and it eventually inspired researchers in the Social Dynamics Lab to study how well the red-blue divide correlates to seemingly nonpolitical choices like what hot beverage to drink.

LIBERALS	CONSERVATIVES
MOST FOLLOWED:	
VEHICLE	
Prius	Harley Davidson "hog"
	
SPORT	
Soccer	Football
	
TV SHOW	
"Real Time With Bill Maher"	"The Walking Dead"
	
MUSIC	
Jazz and R&B	Christian and Country/Western
	



To see where the Twitter feeds you follow fall on the political spectrum, go to the lab's interactive website lifestyle-politics.com (seen at left).

by parsing their tweets hour by hour. "We just couldn't do these things before, because social life is really hard to observe—it's fleeting," says Michael Macy, the Goldwin Smith Professor of Arts and Sciences and the lab's director/founder, who has appointments in sociology and information science. "It's hard to be at the right place at the right time, to see things when they happen. A lot of it happens in private. You just can't send enough people out into the field to track all this stuff down."

As Macy explains, the advent of the Internet and social media isn't just important to sociology because of the vast amounts of data it generates—and in fact, he doesn't favor the term "big data" to describe it, because its value lies not just in its quantity but in its quality. "These digital traces from social media provide detailed, time-stamped indicators of what people are thinking and doing at a granular level, and on a global scale," he says. "It's pretty remarkable." Tweets are inherently public—the modern equivalent of sounding off in the town square—and Twitter makes its database available in a research-friendly format. Reddit data is similarly accessible, as is that of public group accounts on Facebook and Instagram. With the appropriate oversight from the University's human subjects committee and the cooperation of the relevant companies, SDL researchers have also accessed anonymized data on one-on-one communications; this does not comprise actual content, but includes such information as logs of e-mail traffic (from Yahoo!) and phone call patterns (from British Telecom). "The field of computational social science has really taken off in the past few years," says sociologist Duncan Watts, PhD '97, noting that the discipline will have its third annual international conference this summer, in Germany. "It's generating a lot of excitement, particularly among younger scholars, and Cornell is one of the best places in the world for this type of work."

'These digital traces from social media provide detailed, time-stamped indicators of what people are thinking and doing at a granular level, and on a global scale,' Macy says. 'It's pretty remarkable.'

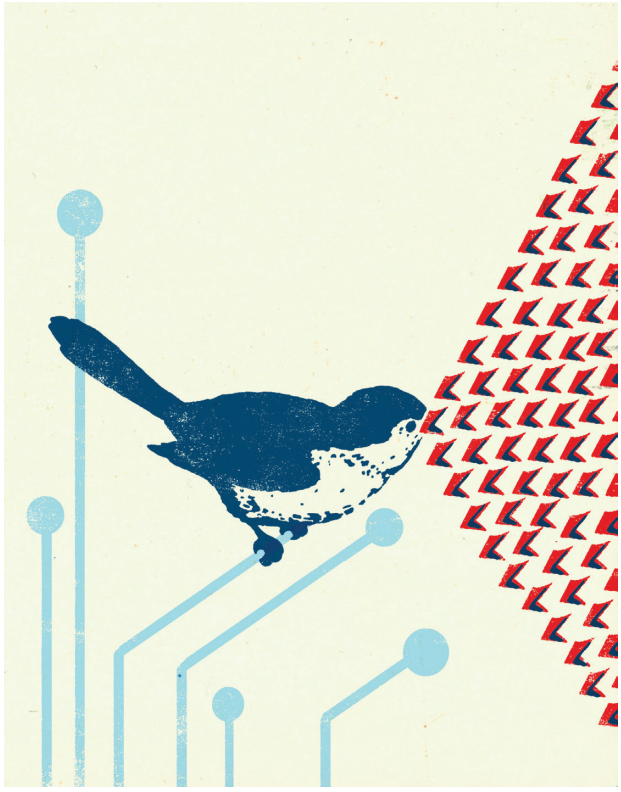
An early pioneer in the field—he did groundbreaking research on the "six degrees of separation" problem as a grad student under applied math professor Steve Strogatz—Watts is now at Microsoft's research division. Something of a legend among Macy's students, Watts has spoken to the lab numerous times and visited campus in his role as an A.D. White Professor at Large; Macy also teaches Watts's 2011 book *Everything Is Obvious* in his popular course on social science prose, Six Pretty Good Books. As Watts notes in it: "Just as the invention of the telescope revolutionized the study of the heavens, so too by rendering the unmeasurable measurable, the technological revolution in mobile, Web, and Internet communications has the potential to revolutionize our understanding of ourselves and how we interact . . . Three hundred years after Alexander Pope argued that the proper study of mankind should lie not in the heavens but in ourselves, we have finally found our telescope."

Watts points out that when he and Strogatz sought a quarter-century ago to establish whether everyone on the planet was indeed connected by six or fewer other people, "the obvious way to answer that question was to construct the network of the world and count how many links there were." And as he notes with a laugh: "We thought about that for thirty seconds and realized it was impossible—it could never be done, it was inconceivable, there's no way it would ever happen, and we had to think our way around the problem." But lo and behold, today such a network exists and is available for study. "Now we have Facebook," he says. "There actually is a network of almost a couple of billion people, and you really can count the links between everyone and everyone else—and it turns out to be less than six. So the brute force approach works after all, but only because this super hard problem that we couldn't even imagine solving got solved through this unexpected route of the Internet leading to the Web leading to social networking sites leading to networks of that scale."

NETWORK ANALYSTS: Sociologists Duncan Watts, PhD '97 (above), and Michael Macy (top).

The trove of online information allows researchers to go where they've never gone before. They can travel back in time—virtually speaking—to witness the formation of a community or movement; they can track the spread of a piece of information, trend, or other cultural phenomenon from its genesis. And while the ›

PHOTOS: MACY, ROBYN WISHNA; WATTS, ROBERT BARKER/CORNELL MARKETING GROUP



Going Viral?

A story, video, meme, or hashtag that becomes wildly popular is said to "go viral." But does it really? As Watts points out, "All the metaphors that we use to talk about social contagion come from biological contagion; they have an underlying assumption that things spread through a multigenerational branching process, where I infect a few of my friends and each one of them infects a few of theirs, and this grows exponentially to infect a large number of people."

That is indeed how diseases like Ebola spread. But at Microsoft Research, Watts and colleagues examined more than a billion tweets, drilling down to ones that got at least 100 retweets. While rare—about one in 3,000—they still numbered in the hundreds of thousands. "We were able to look systematically at the structure of these big cascades and found that stuff doesn't really go viral at all," he says. "Most things don't spread—and even the things that do spread do so mostly because they get retweeted by some big hub, what we used to call the 'Justin Bieber effect.' " For something to become hugely popular, in other words, it must generally be propelled by a celebrity or other super-tweeter with millions of followers. "This should change your intuition about how things spread on social networks vis-à-vis biological networks," Watts says. "But the other thing that's interesting is that you couldn't have done this study if you didn't have all these tweets to start off with. If you're studying rare events, you typically only have one of them—and in this study we had hundreds of thousands. That's the kind of study that just wouldn't have been possible a decade or so ago."

Demographic Data

For many people, the relative anonymity of the Web is one of its attractions, but that lack of data can be a bane to researchers. SDL has been working on ways to parse demographics from tweets and other online data—and those online clues can be even more revelatory than standard self-reporting. Not only do they sidestep the chance that a person won't answer honestly, but they can offer a more accurate picture: a rich vocabulary, for example, indicates a level of intelligence for which education was always just a proxy. "All this data created as a byproduct of people's online interactions and use of things like smartphones give us an unprecedented window into social life at a fine-grained scale that wasn't possible with survey research," says third-year grad student Tom Davidson, whose dissertation work has included studying political polarization during the run-up to the Brexit vote in the UK.

Some of those tactics:

— Education levels can be divined by tallying the vocabulary words used in posts.
— First and last names can inform gender and ethnicity.
— Age, race, and gender can be guessed through facial recognition analysis of profile photos.
— If users have "location services" turned on, researchers can track where they spend their nights and weekends and extrapolate what neighborhood they live in, then use GPS data to identify average home value via sites like Zillow—offering a proxy for income and net worth.

ILLUSTRATIONS: TOP LEFT, RICHARD MIA/THE ISPOT; BOTTOM RIGHT, FILO/ISTOCK.

often-anonymous nature of online communication can be an impediment, investigators have devised ways to tease out information on individuals, from net worth to education levels. "This is a really exciting time," says Milena Tsvetkova, PhD '15, a lab alum now at the London School of Economics. "A lot of people are jumping into social science in general; physicists, computer scientists, mathematicians are now doing social research, applying their skills to these enormous amounts of data online and looking at social problems. It's the best time to be a researcher looking into this area."

What's more, the Internet has enabled experiments on a much larger scale, using platforms like Amazon's Mechanical Turk, a virtual marketplace for piecework labor. "By recruiting online you have this vast participant pool," says Macy. "Plus, it's a much better cross section of the population: you're not just picking up the idiosyncrasies of college sophomores taking intro psych, who've just read about the thing you're studying." As information science grad student Wei Dong, MS '16, notes, such traditional campus-based surveys skew toward subjects who are well-educated and at least middle class. "It misses a lot of people who are not necessarily going into universities or don't speak English," says Dong, whose dissertation work includes analyzing cross-cultural social networks within a large corporation and trying to predict which hashtags will go viral on Twitter. "So we're widening our scope a lot."

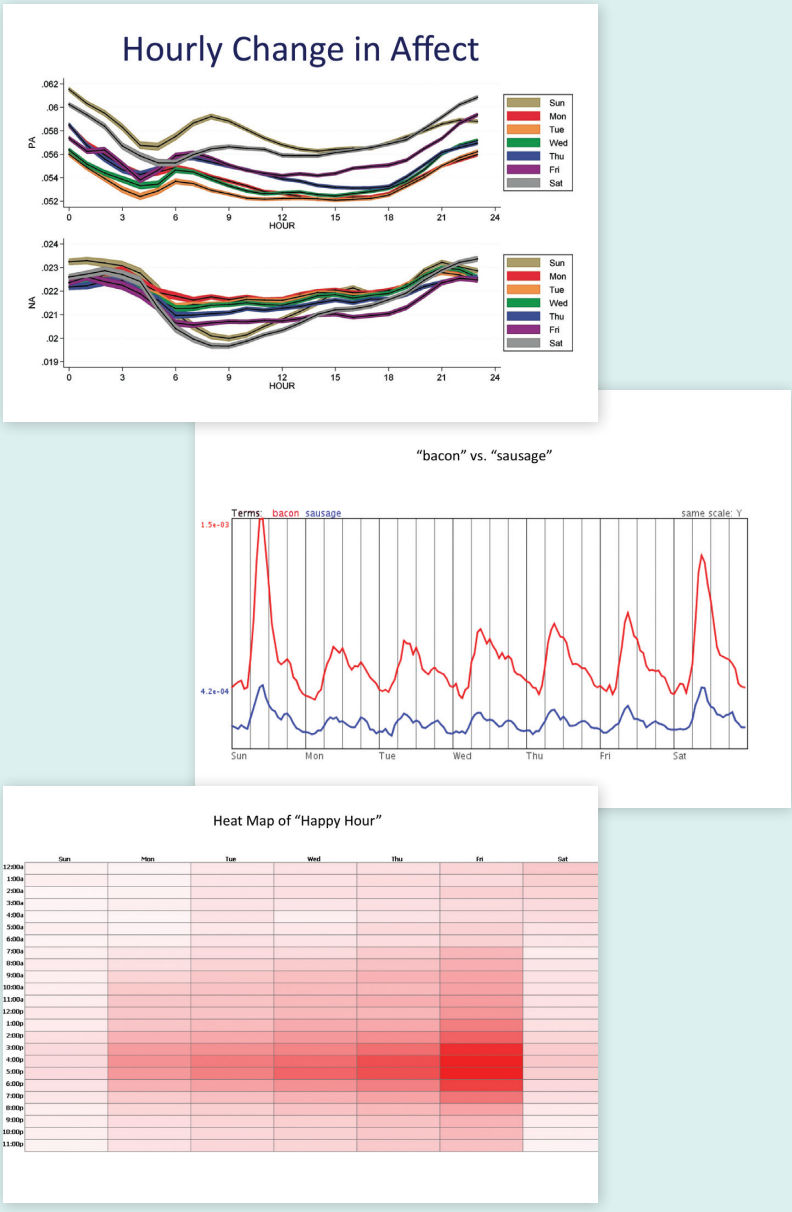
'A lot of people are jumping into social science in general,' says Milena Tsvetkova, PhD '15. 'Physicists, computer scientists, mathematicians are now doing social research, applying their skills to these enormous amounts of data online and looking at social problems.'

Macy stresses that one of the biggest advantages of the social media revolution is that it facilitates not just the study of the individual—what he terms an "atomistic" view of the world—but the larger networks that influence us. This, he says, can offset the perils of so-called streetlamp bias, which he describes as "the tendency to look where the light is pointing, not where the answer lies"—in other words, to seek explanations where you already have data, such as that gathered through traditional survey methods. "The data from surveys is all about the individual," he explains. "It's your education, gender, race, income, religion, age. That's what we use to explain things; we say that your behavior, political opinions, cultural choices, and lifestyle preferences are somehow shaped by these individual characteristics—forgetting that all of that is affected by your friends, neighbors, and other people who influence you."

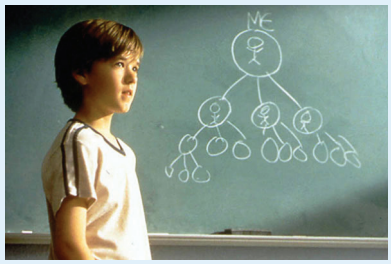
Since the SDL was established a decade ago, it has fostered research on a variety of topics, from hate speech to how information spreads online to the effects of identity-reinforcing "echo chambers," in which people tend to communicate with others who are like-minded; one team of undergrads is currently trying to figure out if "fake news" can be identified not by its content, but by the network of people and groups that link to the stories. Highly interdisciplinary, the lab brings together grad students and undergrads from numerous fields—not only sociology and information science but psychology, applied mathematics, computer science, economics, and more. "It's really great—people are excited to learn from others with different backgrounds, and they ask interesting questions," says George Berry, MA '16, a fifth-year grad student in sociology who's using Twitter data to study patterns of interaction on social media across lines of gender, race, and income. "There's active sharing of information about a new method or piece of research, and we frequently workshop each other's papers. Professor Macy does a really good job of fostering that. He's always looking for people who can bring a fresh perspective, teach us something new, or help us understand something that might be difficult. I've learned a ton from that environment." ›

Rhythms of Life

To track moods using Twitter, Macy and colleagues studied 530 million messages by 2.4 million users worldwide. They divided the tweets into "buckets" representing the 168 hours of each week, then counted the words that denoted positive affect (hilarious, fantastic, awesome) or negative affect (anxious, embarrassed, depressed). The upshot: mood peaks early in the day and more or less goes downhill until evening—regardless of day of the week, season, or culture. The researchers also parsed how often people tweeted words like "bacon" and "sausage"—they found, for example, that the former is a cherished weekend leisure food—and even showed that the idea of a "happy hour" is literal: people are jolliest on Friday afternoon and early evening. The work was published in *Science* in 2011, with grad student Scott Golder as lead author.



WITH FEELING: The top illustration depicts how mood fluctuates during the day, color-coded by day of the week; tweeted words denoting positive affect are shown in the first graph, negative affect in the second. Middle: The incidence of the words "bacon" (red) and "sausage" (blue) in tweets. Bottom: The frequency of tweeted words denoting happiness, which is strongest on Friday afternoons.



LITTLE HELPER: Haley Joel Osment as a boy who does good deeds in the movie *Pay it Forward*.

The Greater Good

The 2000 film *Pay It Forward* explored the idea of how spontaneous generosity can spread in an ongoing chain. Macy's lab decided to investigate the phenomenon with two large-scale online experiments. "We tested the hypothesis that when a stranger does a favor for you, you become more likely to do a favor for another stranger—and they become more likely, and it ripples down, a cascading effect of good behavior," he says. "But we went beyond the movie, and looked to see if bad behavior also has this cascading process."

With funding from the NSF, researchers recruited nearly 2,000 people via Amazon's Mechanical Turk platform. Each player got a stipend of actual money; in the first experiment they could share some of it, while in the second they could steal from others. Tsvetkova and Macy (who went on to co-author an op-ed in the *New York Times* about the study) found that just observing largesse didn't necessarily make a person more generous—evidence of the "bystander effect," in which we assume that if others are helping out, we don't need to—but actually benefitting from generosity did make them more likely to pay it forward.

On the flip side: not only did being stolen from make players more willing to do a bad turn themselves, but just *observing* antisocial behavior made them more likely to do so—evidence of the so-called "broken windows effect," which holds that a negative environment engenders misdeeds.

COLLABORATIVE RESEARCH: Macy in the lab with grad student Dana Warmsley, MS '15 (left), and undergrad Pujaa Rajan '17 (right).

Berry's work, basic research funded by the National Science Foundation, is aimed at understanding whether the Internet promotes communication across disparate groups. "We always hear about political polarization and income inequality—and if the Internet is facilitating interaction across class boundaries, that's really important to know," he explains. "And if it's not, we want to know that, so we can think about how we can better facilitate conversations among everyone." This fall, Berry will become the latest alum of the lab to join Facebook's Core Data Science team—what Macy calls "the Bell Labs of social science"—where he has interned the past two summers. In early April, he went to Australia to present a paper based on research he conducted at Facebook on how the quality of comments within a discussion—essentially, whether they're civil discourse or all-caps rants—can affect the tenor of subsequent postings. "Sociology has a huge number of compelling perspectives on a lot of different aspects of social life," he observes, "but for a long time our discipline has been in the position of wanting to ask really big questions that we didn't necessarily have the tools to answer."

Pujaa Rajan '17, one of SDL's undergraduate researchers, came to the lab via a circuitous route; the Nebraskan started out as a math major, then switched to computer science, but found it too theoretical. She settled on information science, and is currently doing an honors thesis on her work under Macy, including her contributions to an ongoing project using Twitter to see how cultural preferences are polarized based on political identity. "What's really special about the lab is that it combines the computational and quantitative aspects of looking at data with the social aspect of it," she says. "This lab is really good at using technology, the newest coding algorithms and all that, to research how people interact with each other. You're not just trying to discover a new math formula—you're using those methods to come to conclusions about the way people live their lives." ■

PHOTO: ROBYN WISHNA

Battling Hate Speech

At SDL, online hate speech—and how to discourage it—is a hot topic. Doctoral student Dana Warmsley, MS '15, is devoting her dissertation to the subject. "Especially now, with the recent climate of hate speech in American politics, we thought this was really interesting," says Warmsley, who majored in math at CUNY's Hunter College. "We're hoping to illuminate the hidden population of hate speakers, those who aren't part of a big group or organization."

Collaborating with Davidson, Warmsley has been trying to find efficient ways to identify hate speech on Twitter using keywords—but that can be more complicated than it sounds. "The definition of hate speech has to do with the intention to humiliate, degrade, and even threaten people based on their characteristics, whether it be age, race, gender, things like that," Warmsley explains. But as Davidson points out, "Particular words are often used in different ways, so even those that sound sexist or racist might have different connotations. For example, someone quoting lyrics from a rap song is using that language differently from a white supremacist."

The researchers tried to account for such ambiguity by having volunteers read tweets and decide whether they contain hate speech or simply offensive language. "After having people evaluate tens of thousands of tweets, we were able to train a machine learning model to use this data to accurately identify hate speech in other tweets," Davidson says, "many more than we would be able to get people to actually read." The team is also analyzing the demographics of hate speakers—"age, race, gender, political affiliation, education level, even personality traits," Warmsley says, and studying their networks: "how they interact with other users on Twitter; are they connected to other hate speakers, or just to the average Joe? How are they spreading hate, and what kind? Who are their targets?"

One aim of this kind of work, Warmsley says, is for social networks like Twitter and Facebook to efficiently and accurately identify hate speech by their users. "You want to be able to detect when people are being harassed," she says. "Given all the tweets, you can imagine how long it would take for humans to go through them and take the proper actions."