The Echo
of Digital Tweets

Twitter, Facebook and their ilk – social media are increasingly dominating the Internet. But how do messages spread across these new platforms? What role does a small clique of super-influentials play? And to what extent are the traditional mass media also leading the pack online? These are the questions that interest **Krishna Gummadi** at the **Max Planck Institute for Software Systems** in Saarbrücken.

TEXT **RALF GRÖTKER**

Krishna Gummadi has no fewer than 1.75 billion tweets – text messages from the social media service Twitter. The company allows its users to "tweet" messages free of charge from any computer or Internet-capable cell phone to other users who have become "followers" of the sender. Tweets are limited to a maximum of 140 characters (which corresponds to the length of this sentence), usually including a link to a website. "A goldmine," says the 31-year-old Indian. The treasure is securely stored on 58 computer servers in the "Wartburg," an imposing parish hall built in Saarbrücken's town center in the 1920s.

Here, in the immediate vicinity of the bank and a credit card company, is the home of the Max Planck Institute for Software Systems, at least until the expanding institute's new building on the university campus is finished. Gummadi has headed the research group "Networked Systems Research" since 2005. To understand his passion for the 1.75 billion tweets, we have to go back a bit further.

In early 2003, the SARS epidemic broke out at the Prince of Wales Hospital in Hong Kong. Investigations later showed that a single patient had directly infected 50 other patients, which led, in the end, to 156 SARS cases in that hospital alone – and then to the outbreak of the epidemic well beyond the city.

DO VIRUSES SPREAD LIKE FASHION TRENDS?

It seems that ideas and fashion behave in much the same way as diseases. The sudden success of the Hush Puppies brand is one example of this. In the mid-1990s, sales of this comfortable crepe-soled footwear had reached an all-time low. Then, suddenly, the unexpected happened: New York fashion designer John Bartlett ordered a series of Hush Puppies for the presentation of his spring collection. The shoes had come to his attention because some people in New York's club scene had begun to wear them. A Hush Puppies epidemic broke out. In 1995, the company sold 430,000 pairs of shoes – 400,000 more than in the previous year. The following year even saw nearly two million pairs sold.

The American science writer Malcolm Gladwell, who describes the story in his book *The Tipping Point*, has a simple but plausible explanation for such occurrences. Epidemics are triggered by influentials – people in a particular professional and social position, but also with a certain talent and attitude toward life that allows them to come into personal contact with a vast number of people.

The social epidemic theory that has been circulating as standard knowledge since as far back as the 1950s, es-

The bird tweeting the latest news from the rooftop is blue and serves Twitter as its logo (left). Users use photos or icons as identifiers (right).

···

pecially in the world of marketing, has been sharply criticized again and again in the more recent past. One objection is that the spread of viruses and fashion can't be compared because, in the case of a virus, the risk of infection upon repeated contact with the pathogen is always the same, but with a fashion trend, both keeping-up and desensitization effects can occur. It was also criticized that people like Gladwell were choosing anecdotes specifically to suit their purposes.

A closer look at the events often does, in fact, show that those who are presumed to be key people are merely a product of the circumstances. Particularly the above-mentioned SARS outbreak in Hong Kong is a perfect example of this. In the Prince of Wales Hospital, everything began when the patient at issue was incorrectly diagnosed with pneumonia. Instead of isolating him, he was put in a crowded, open hall with poor air circulation. He

was hooked up to a lung ventilator – which spread the SARS viruses around the area. The case is thus a very poor example to serve as proof of the influentials theory.

## EMPIRICAL STUDIES OF SOCIAL NETWORKS

This is where Krishna Gummadi comes in. The research he conducted primarily with his colleague Meeyoung Cha, who now works as an assistant professor at the Korean Institute of Science and Technology, promises, namely, to decide the dispute over the influentials. For the past several years, Gummadi has been examining the information flows of social networks in Internet communities such as Facebook, LiveJournal, LinkedIn (the American counterpart of Xing) and Twitter.

What happens within these online networks, and how, for example, trends spread, also provides important clues

about the spread of viruses and social epidemics in the physical world. After all, the fundamental structures in both cases are networks – and these definitely do have comparable properties.

Gummadi's group published some of the very first works to trace the development of online networks on a large scale in 2007. Most recently, he has been studying the usage behavior of the online service Twitter in a major series of studies. In the second half of 2009, the researchers "crawled" information – with special permission from Twitter and taking the usual data protection guidelines into account – from nearly 55 million Twitter accounts. Information that is, in principle, publicly available – just not in
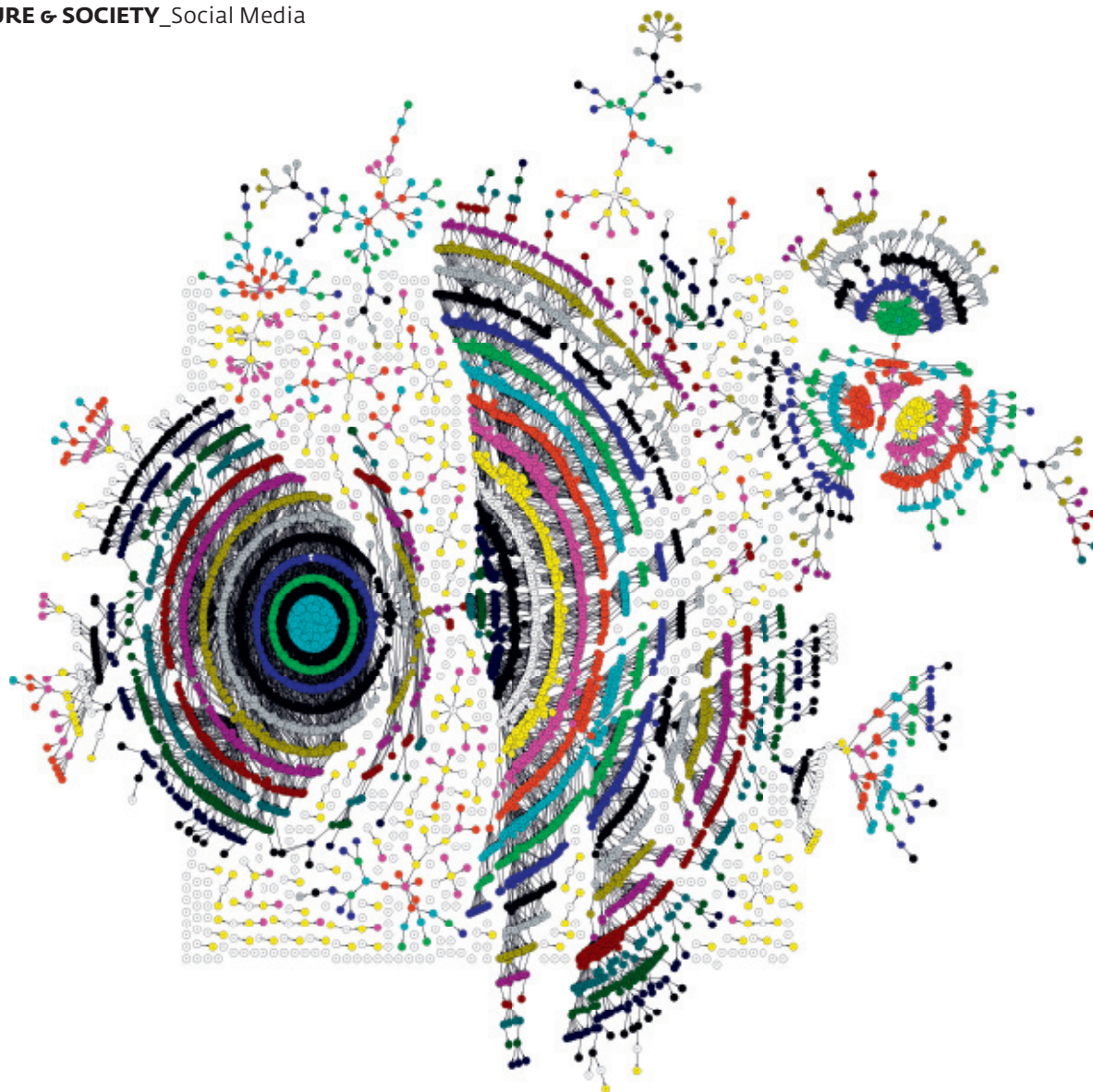
A message makes waves: The graphic shows how a web address is circulated through Twitter. Each of the more than 14,000 nodes stacked in the large circulation trees, also on multiple levels, represents one user. The colors reflect the level of circulation (white, yellow, and so on – the color coding repeats after the 15th circulation level). The starting point and the first extensive node of the prominent tree in the left half of the image are hidden under the turquoise dot.

collected form – was gathered through repeated data queries to the Twitter website. From these 55 million, another 6 million active users were filtered out. These users, in turn, had sent the aforementioned 1.75 billion tweets.

Using the new data, it is now possible for the first time to examine empirically how ideas and trends spread in social networks – and in this way to sound out popular theories on word-of-mouth marketing or the diffusion of innovation. One study on which the Max Planck team is currently working is aimed at showing whether it is different users or always the same ones who successfully point their Twitter colleagues to a certain website. Until very recently, such questions could be examined only with the aid of computer simulations or other models.

RANKING OF INFLUENTIAL TWITTER USERS

Duncan Watts, formerly a sociologist at New York's Columbia University and now in the research department at Yahoo, concluded from a test of various modeling methods that the influentials could, under certain circumstances, play a role – but that these circumstances are very narrowly defined and presumably occur relatively infrequently. Influentials, his study concludes, are not so much people who have certain characteristics as those who have simply gotten lucky.

And this is precisely where the new Twitter study picks up. In formulating hypotheses and designing experiments, the study relies on established methods from the field of network analysis. The approach, however, is empirical. The researchers systematically combed the data to find out which websites the Twitter users had recommended to their followers. Then they compiled a ranking aimed at showing which users had contributed the most to spreading a certain web-

> » Nothing happens without the mass media. Nevertheless, the great majority of less well-linked Twitter users act as an amplifier for the messages posted by the mass media.

site. To this end, the researchers trawled through millions of Twitter posts for any mention of a website address, and then examined these, in turn, for correlations and patterns. "It was a pretty hairy business," says Gummadi. Tasks like this bring even large computer systems to the limits of their capacities.

An obvious objection here is that the behavior of online networks like Twitter implies just as little about other social phenomena as the spread of viruses does about social epidemics. In a strict sense, this is true. On the other hand, it is precisely by analyzing social media that one can develop a sense of which subtleties can make a key difference – both online and on the street.

For example, the study of a team led by Jure Leskovec from Carnegie Mellon University showed that most recommendation chains on the popular US product recommendation site Epinion end in a gaping void after a few links – which doesn't exactly bode well for the spread of social epidemics. The photo site Flickr, which was the subject of a study that preceded Gummadi's Twitter analysis, is a different story. "Most images do not spread far here," is the gist of Krishna Gummadi's findings. "A few superstars among the photos account for the majority of recommendations." So the conditions here are more favorable for epidemics.

A recently published study based on the Twitter data examines in detail how the various communication paths within Twitter differ. On Twitter, it is possible to simply "follow" someone – that is, to subscribe to his or her tweets – without this following needing to be confirmed by the other person, unlike "friending" on Facebook. Further, one can retweet messages – that is, forward messages one receives – to one's "fol-

lowers." And finally, one can reply to messages from any Twitter user, which results in a "mention" for that user.

CELEBRITIES RECEIVE THE MOST MENTIONS

If we examine these three forms of action independently, we no longer have one, but essentially three separate Twitter networks. As a result, we find that, in the category "users with the most followers," news sites, celebrities and politicians were the most influential players. The news sources likewise led in the category "retweeted," along with prominent business advisers. Finally, the celebrities received the most "mentions." This means that having thousands of followers does not necessarily make a Twitter user an influential – or, in any case, it is no indication that the message one wants to get out is actually picked up and circulated.

The study also showed that influential Twitter users are usually successful in a variety of fields, not just on a single topic (unlike, for instance, the communities on the Epinion platform, which are usually about a specific product). Also, the influentials are not just one-hit wonders. They remain successful over an extended period. "Influence here is not a product of chance, but the result of concerted efforts," finds Gummadi's study, countering Watts, the influentials skeptic.

And something else is striking: the prominent role that mass media and their representatives play in online social networks. This role does not necessarily fit with how we conceive of word-of-mouth marketing. Krishna Gummadi's team took this observation as an occasion to launch a new study. "We wanted to know what happens when we simply remove highly linked players from the game, like the

Tracking a message: Massimiliano Marcon, Krishna Gummadi, Bimal Viswanath and Nuno Santos (from left) discuss how information spreads in social groups.

Photo: MPI for Software Systems – Bimal Viswanath

sites of news broadcasters and newspapers," says the network researcher.

They selected the emergence of British amateur pop singer Susan Boyd in Twitter posts in 2009 as a test event. The outcome: 60 percent of all Twitter users first learned of Boyd from tweets posted by news sites. Meaning: Nothing happens without the mass media. Nevertheless, the great majority of less well-linked Twitter users acts as an amplifier for the messages posted by the mass media. After all, this majority contributed 5 percent to the spread of the news about Susan Boyd.

**USEFUL ANALYSES
FOR MARKETING EXPERTS**

There are many other things that could be investigated using Twitter data. "I would be interested in knowing whether services that are used for recommendations lead to an increase in the spectrum of individually consumed media – that is, whether I read things that I would never have known from traditional mass media channels," says Gummadi. Another project: "We could find out how fashion trends and customs spread." For instance, it has recently become common on Twitter to use so-called tiny URLs instead of mile-long website addresses. The

data set would permit highly precise tracking of who brought what type of tiny URL into play and when. "Something like this would be useful as a starting point for finding out how innovations spread."

The results of such analyses could definitely be useful for marketing experts. Here, too, it is now thought that the power of influentials plays a lesser role than was previously believed. "The Web has changed. Today, nearly everyone is registered on Xing or Facebook – it is hardly possible anymore to say who is the most relevant," says Christian Wilfer, managing director of Dialog Solutions, an agency that specializes in viral marketing. Further, "It always depends on the product," adds word-of-mouth expert Martin Oetting of trnd. "If I'm advertising a fabric softener, there is hardly likely to be a group of particularly influential communicators. Normal consumers who are highly involved are far more important."

The trnd marketing community devised its strategy accordingly: instead of just a few consumers, it "immunizes" many thousand who have registered as product testers on trnd's online platform with the latest products. But unlike Dialog Solutions, which pursues a similar plan with its Share-

ifyoulike platform, trnd concentrates on offline communication – because after all, people listen to their friends differently than they do to "friends" on Facebook.

Analyzing networks in social media in the manner in which Krishna Gummadi and his colleagues in Saarbrücken do is relevant for practical applications, but not just for marketing and innovation research. Knowing exactly how networks are structured is also a requirement for algorithms for fighting spam. Spammers boost their online reputation by linking their sites and user accounts as much as possible. But this approach often results merely in closed link universes – with no connection to the clusters of honorable netizens. Algorithms can aid in identifying and crippling such spam clusters.

**SOCIAL SEARCH POSES
COMPETITION FOR GOOGLE**

Another area of application is social search: searching, not in the World Wide Web, but within the community of like-minded individuals. A prototype of such a search method was developed a few years ago under the direction of Alan Mislove, a former student of Gummadi's who is now an assistant professor at Northeastern

The daily emotional roller coaster in the US: Based on the words used in Twitter messages, with points accorded to certain mood-related words, it is possible to assess the mood of Twitter users in the individual US states at various times of day. Red indicates a rather bad mood, green a better one. According to this information, people in the western and southeastern states are generally happier, but the mood hits a low everywhere in the early afternoon. The states are distorted because their surface areas indicate the share of messages sent from there in all tweets in the US.

Graphic: Alan Mislove

Twittering during the 2010 soccer World Cup: The graphic indicates, for the period from June 11 to July 11, how many Twitter messages per second contained, as a marked keyword, the name of the country whose flag appears in the background of the peaks. To better distinguish the message density of the selected countries, some of the peaks point upward and some downward. Overall, the number of tweets with country names rose toward the end of the WM. The total number of tweets per second is shown in gray in the background.

University in Boston. "We wanted to know how much one can benefit from searching in their social environment," says Mislove.

Whenever one of the ten researchers involved in the project called up a website, its description and contents were stored in the institute's internal network. If one of the researchers then searched the Internet, he or she would see, in addition to the Google list, a list of the websites the team had visited. The advantage here is that "PeerSpective," as the project was called, also showed entries in online library catalogs that aren't listed in Google.

The pilot study showed that nearly 8 percent of all of the search results that were actually shown could be found with PeerSpective alone. It is seldom that anyone stacks up so well against Google. Mislove succeeded in doing this with software that he cobbled together by himself in just one week.

The first search engines using the principle of PeerSpective are already on the market. One example is Swicki, from the California-based company Eurekster. The research departments of Yahoo, Google and Microsoft also have plans for social search. However, novel search engines have not yet taken hold as strongly as integrated solutions along the lines of "See what your friends are sharing on Facebook" and "E-mailed – Blogged – Viewed most" on the pages of the New York Times, or the integration of "Buzz" in Google's "gmail."

Regardless of what approach is taken to social search, in order for it to work, one first has to find the right community for a given search query. Gummadi and his team recently addressed this issue. They tried to use one of the usual algorithms to detect groups within a network in order to filter people with similar interests. Sur-

FISHING FOR INFORMATION

Much of the content of social media platforms is publicly accessible because the users have agreed to its publication. Normally, however, only the company operating a social media platform can access all of this data to, for example, analyze it for statistical purposes. But it is also possible for outsiders to use automated search queries to gather and pool information that is otherwise scattered throughout a given social network. These search queries are done with digital search robots, also known as "web crawlers." Internet search engines also use crawlers to create an index of the sites available on the World Wide Web. This index is then accessed when an actual search query is run. Most social media sites limit the number of search queries that a given Internet user may execute, so that third parties can't benefit economically from the data through database analyses. Researchers who want to "crawl" these sites to obtain information about the nature of network structures in social media therefore usually have to request permission from the social media service in question.

Who tweets about the Max Planck Society and when? For the week from November 23 to December 1, 2010, the graph shows when messages were sent containing the words "Max Planck". Symbols stacked vertically indicate tweets that were sent at the same time. Tweets appear closer to the center the more replies they received or the more frequently they were retweeted. Messages that appear large are either new or – as here, in the case of a retrospective analysis – randomly selected messages. The Max Planck Society has been active on Twitter since October 2010.

prisingly, however, the algorithm did not allow them to detect a few groups that they knew existed.

TWITTER MESSAGES AS A PUBLIC OPINION BAROMETER

First, a bit of sociology was needed here. It turns out that not all groups are created equal. Some groups (called "communities" in sociology) are held together by personal connections: "Just like a few of us here at the institute meet each week to play poker – because we simply like to hang out together," as Krishna Gummadi says with a smile. Other groups (the "societies") are subject-based – "like Greenpeace." In subject-based groups, most of the participants don't know each other at all. Gummadi remarks, "We first had to become aware of this difference. That is why the algorithm we used first, which merely filtered participants that were connected through particularly few links, couldn't find these clusters."

Listening to Krishna Gummadi, one gets the impression that the range of practical issues to which the results of network analysis could have a di-

rect bearing on social media is nearly endless. One of the latest projects is a national public opinion barometer. Alan Mislove analyzed Twitter messages in the US with a view to what emotional state they reflect. His findings are displayed in an animated map in which the individual states are ranked on a color scale from green (happy) to yellow (neutral) to red (not happy) – and change color over the course of the day.

By itself, the public opinion barometer is just fun and games. But it demonstrates what information lies dormant in social media. The operators of Twitter know literally how the world ticks. They can watch what people are talking about and what they are doing with their money. From this, one can derive forecasts – whether share prices of automotive companies will drop, or how much money a newly released film will take in at the box office by the end of the first week. That itself is fascinating. On the other hand: "It's really crazy how much power these companies have," says Gummadi. So it is understandable why someone would consider 1.75 billion tweets to be a goldmine. ◄

GLOSSARY

Social media
Online platforms for sharing opinions, messages and media content, such as photos, within social networks. Popular social media sites in Germany include the business community Xing, the StudiVZ platform, aimed at students, Facebook, the photo site Flickr, and Twitter.

Word-of-mouth marketing
Spreading recommendations or information through personal communication. Network structures, where, similar to the route network of international air traffic, a few hubs act as distributors, promote the quick spread of advertising messages. A special form of "word of mouth" is collective filtering, which is used, for example, for the personal book and media recommendations at Amazon.

Social epidemic theory
According to the explanatory model that has recently been supported particularly by American science writer Malcolm Gladwell, opinions and advertising messages spread in a similar way as pathogens. The influentials theory plays a key role in this model. According to this theory, the development and magnitude of an epidemic are driven primarily by the existence of a small group of people who have a particularly large number of social contacts.

Graphic: Moritz Stefaner