DiVA – Digitala Vetenskapliga Arkivet <http://umu.diva-portal.org>

This is an author produced version of a paper published in **Journal of Computer Science and Technology**.

This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the published paper:

Holme, Petter; Huss, Mikael

Understanding and Exploiting Information Spreading and Integrating Technologies

Journal of Computer Science and Technology, 2011, Vol.26, Issue: 5, 829-836 pp

<http://dx.doi.org/10.1007/s11390-011-0182-3>

Access to the published version may require subscription. Published with permission from:

Springer

The original publication is available at www.springerlink.com

Understanding and exploiting information spreading and integrating information technologies

Petter Holme^{1,2} and Mikael Huss^{3,4}

¹ IceLab, Department of Physics, Umeå University, 901 87 Umeå, Sweden

² Department of Energy Science, Sungkyunkwan University, Suwon 440–764, Korea

³ Science of Life Laboratory Stockholm, Karolinska Institute Science Park, Box 1031, 171 21 Solna, Sweden

⁴ Department of Biochemistry and Biophysics, The Arrhenius Laboratories for Natural Sciences, Stockholm University, 106 91 Stockholm, Sweden

E-mail: petter.holme@physics.umu.se; mikael.huss@scilifelab.se

Abstract Our daily life leaves an increasing amount of digital traces, footprints that are improving our lives. Data-mining tools, like recommender systems, convert these traces to information for aiding decisions in an ever-increasing number of areas in our lives. The feedback loop from what we do, to the information this produces, to decisions what to do next, will likely be an increasingly important factor in human behavior on all levels from individuals to societies. In this essay, we review some effects of this feedback and discuss how to understand and exploit them beyond mapping them on more well-understood phenomena. We take examples from models of spreading phenomena in social media to argue that analogies can be deceptive, instead we need to fresh approaches to the new types of data, something we exemplify with promising applications in medicine.

Keywords Modeling and prediction, Computer-supported collaborative work, Health care

1 Background

The idea of an information age replacing the industrial age has been around for about forty years [1]. In this new era an incessant flow of information permeates all levels of society, too much for any individual to process. Those who can filter out the important bits, however, will have an edge as important as owning resources was the industrial age. The other way around—if one cannot process all the information one will suffer from overload of it (a maxim that still can be heard, even if it was more common in the 70's and 80's, cf. Ref. [2]). With the information age comes a new individual-

ism where media is no longer one-to-many mass media but personalized to the interests of the individual [3]. It took until five or so years ago until many of these ideas to finally became reality. Social media (Twitter, Facebook, etc.) replace some of the roles of telephone and face-to-face meetings [4], recommender systems replace hours of work browsing catalogues, etc. Today we have the full feedback loop, from the data we generate to the output of these technologies and back to our behavior. To understanding and utilizing social media are thus both parts of the eigenvector equation of

our future society.

For many people in the computerized world, producing and using digital information is both their job and pastime. This information-centric lifestyle that has developed over the last decades has given social science a new universe to observe. Furthermore, developments in statistics give us new telescopes for this exploration. Extracting information from these digital traces are more than an academic endeavor—among the world's largest companies, there are several that mainly focus on aggregating, mining and presenting information. A classic example of such products is search engines and their ranking of webpages matching a query; another example is recommender systems and personalized advertisement. The last few years, companies extracting news and trends from computerized communication, primarily Twitter and similar social media, have emerged. Such information is then fed back to the people and affecting their behavior. There are many open questions about social information spreading. We know fairly well how far messages can spread in e.g. Twitter, but not much of what makes someone spread a message further. This is, we speculate below, reflected in that the terminology is laden with rather mismatched analogies—like “viral videos” that do not spread like (neither human nor computer) viruses. Anyway, a better understanding of information spreading would give better recommender systems and social media. In the rest of the text we will elaborate on some different topics relating how to understand information spreading and how to get better predictions by integrating different data sources and some promising examples of technologies integrating experience and knowledge from people in medical applications.

People get the information they need in their lives from both mass media and their social networks. Lately, social media like Twitter, Facebook and Google Buzz has formed an intermediate channel where people can get information from selected others. This information is typically semi-private in the sense that it is not phrased in way to

be understandable to the public, but it is not directed to anyone (like in private communication like e-mails or face-to-face conversation). There is no sign of social media, as an information channel, losing its importance, which makes business and advertisement strategists looking for possibilities of exploiting it in targeted advertisement and product recommendations [5]. Future recommendation algorithms that integrate information from social media [6,7] would of course benefit from understanding how information propagates in such channels. It is a notoriously hard question since, even if we can monitor actual conversations, we do not know the context of the message and the intent of the person communicating it. This means that before we have thorough psychological studies of such decision processes we have to, to some extent, rely on assumptions and analogies. In this section we argue that we can do better assumptions by experiencing social media ourselves, than by trying to match it, in analogies, to already well-studied phenomena.

Outbreaks of popularity for no obvious reason are a classic social phenomenon in the consumer society. Rubik's cube and Hush Puppies [8] in the West, *nata de coco* and *Winter Sonata* in the East. In more scientific terms, the distribution of economic success of a product is much broader than the distribution of either their perceived inherent quality or their advertisement budgets. The mechanisms behind such fads can be explained in terms of emergent social phenomena—a result of the interaction between large numbers of people. The literature described two types of interaction that can generate such phenomena. The first class of interaction is a threshold behavior of the individuals with respect to social influence—if at least a certain fraction of someone's acquaintances solve Sudoku, then so will that someone [8,9]. The other class of interaction is person-to-person contagion—fads spread like disease. Logically, since it only requires one contact, contagion should be able to spread something faster. Probably it is a better

model than the threshold behavior when it comes to explaining fads in electronic media. It could suffice with just one friend recommending a YouTube video for you to see it. In this type of spreading phenomena, we are acting on impulse rather than trying to conform to a social norm. The idea that items of electronic media—video clips, photos, jokes, and tweets—can follow disease dynamics has even entered the public mind in expressions like “viral video”. Or, maybe scientists have taken this expression from the public as an assumption, perhaps too easily. In this section, we will discuss how far one can push this analogy. Are epidemic models really capturing the mechanisms of viral videos and retweets? Can they even make good predictions?

The, by now, common system to study social-media information spreading is Twitter. This is an electronic media where users post short messages (140 characters or less) about anything. There primary way of reading the messages is to “follow” someone, which in practice means the messages of the person you follow will be listed in your Twitter client or on your homepage at twitter.com. A user can “retweet” other’s tweets of her liking. Chains of retweets thus represent a type of active information spreading, and is currently our best source of empirical data on any type of digital, person-to-person information spreading. First, are there viral tweets—retweet events that spread like a disease? The answer from a recent study [10] is both yes and no. Yes, there are tweets that are retweeted hundred standard deviations more than the average—something like an information outbreak. The probability distribution of the number of retweets is broad and right-skewed, but seemingly not strictly power-law. Even so, judging from Fig. 1, a relationship $P(n) \sim n^{-\gamma}$, where n is the number of retweets of a tweet, and $2.5 < \gamma < 2.7$, would not be too bad a model. But, this exponent would predict a largest observed retweet size of about 24,000—almost ten times larger than the actual value. This suggests a negative feedback restricting n . This would then be a no to the question

if retweets can go viral in the sense of disease spreading models.

In the rest of this section we will make a sketch of how one can model information spreading via retweets and contrast it with models of disease spreading.

2.1 Large-scale observations: Cut-offs in retweet sizes

We take a power-law with a cut-off as our model for the size distribution of retweets. There are a large number of known mechanisms generating power-laws [11], but where does the cut-off come from? We will discuss a first-principle mechanism generating cut-offs (rather a model than disease spreading, diffusion, threshold dynamics or any economic model of information as capital).

A common assumption of epidemic models is that the transmissibility—the probability that a contact between two persons transfer the disease—is independent of the persons in contact. Sometimes models assume a heterogeneous set of agents and different transmissibilities between patients of different types (HIV for example is estimated to be more likely to spread from a man to woman in a vaginal intercourse than from a woman to a man [12]), but never that the transmissibility between the same type of persons can be different in different parts of the network, or different for different times. A common assumption about social systems is that they show assortative mixing—similar people interact more often with each other than with people of different background, ethnicity, interests, political views, etc. This tendency is reflected in social networks—persons close to you are more similar to you than those in more distant parts of the network. Assuming a video becomes popular enough to spread outside of a certain part of the network where people’s interests are different than where it originated, wouldn’t it face a lower transmissibility? This scenario is illustrated in Fig. 2.

To illustrate how one can investigate this sce-

nario, we investigate a Susceptible-Infected-Removed (SIR) model [13] on a stratified network model (i.e. the nodes connect assortatively with respect to a one-dimensional trait). The Susceptible state represents someone might have read the tweet but has not retweeted it, a node getting Infected represent an agent retweeting the tweet, when this tweet gets too old the node turns to the Removed state. The Infected state lasts for a time $\delta = 10$ after which the agent goes to the Removed. This is probably not the best model of retweeting, but it serves our purpose of investigating to what extent one can push the analogy between disease and information spreading. The network model works as follows: assign integer traits between 1 and A (we use $A = 100$) randomly to N nodes and connect a node pair (i, j) with a probability $p \exp(-a \Delta_{ij})$ [14], where Δ_{ij} is the trait difference between i and j . We assume periodic boundaries of the trait space so the difference between a trait 1 and A is 1. We use $p = 0.02$, $a = 2$, $N = 10^4$, giving an average degree of about 2.6. To model the decreasing probability of disease transmission with Δ_{ij} , we let the transmissibility from a vertex to i be $1 / (\exp(-\Delta_{oi} / \alpha) + 1)$, where Δ_{oi} is the trait different between i and the seed. In Fig. 3 we show results for this model.

2.2 Large-scale observations: Broad retweet trees

Another observation from Fig. 1, is that the probability distribution of the generations of retweets, the height of the retweet tree, is rather small—out of 10^7 retweet trees, none was higher than 11 generations. Given only the height h and size N (number of nodes) of the tree, we believe the best measure of the width is the branching ratio that, if all nodes were identical, would give a tree of the height h given the size—i.e. b , where $b^{h+1} / (h + 1) = N$, we call b the average branching ratio (although the word average should be understood from this equation). The largest retweet trees of Fig. 1 have $N \sim 3000$ giving $b \sim 2.4$. Compare this with an epidemic like the swine flu outbreak of 2009 that lived for about 150–200 cycles and in-

fecting about 30 million people (giving an average branching ratio of 1.11–1.16), or the e-mail chain letters of [15] that was found to live up to 288 generations (an average branching ratio of circa 1.05). The large branching ratio of retweets calls for an explanation. In this essay we do not attempt a model-aided explanation but to make a hypothesis. In a blog post analyzing a retweet chain¹ linking to an advertisement at YouTube, the author observes that the celebrity actress Alyssa Milano has been pivotal in the retweet cascade. Media celebrities, typically famous from TV, film and traditional media, naturally have most followers on Twitter. The communication with these hubs is typically one-to many, like traditional mass media. If a retweet cascade passes a hub, it gets relayed to a large audience and the cascade has a chance of becoming large enough to end up in the tail. It seems like Twitter is something between a mass medium and a medium for person-to-person information spreading, and the celebrity hubs are keys in spreading retweets to a large audience in the large retweet trees. In epidemiology it is known that a broad distribution of contact rate (or network degree) speeds up the disease spreading [13]. This is thus maybe an example where an analogy to another system can be helpful, except that the appearance of hubs in Twitter comes from other mechanisms than the power-law degree distribution of some contact networks over which disease can spread (like the network of sexual contacts [16]).

To summarize this comparison of Twitter dynamics and the other systems it has been described as similar to, analogies can help, but only if they come out of a first principle study. Tweets are, in general, not like disease outbreaks, not like fashion trends, not information diffusion, not two-person communication, not mass media, not a social network, not a news media. Twitter is a little bit of

¹ <http://140kit.com/pages/oldspice-research> read September 2010.

everything, and has its unique features, but for a scientific perspective these analogies do not contribute to our understanding of the system—Twitter is Twitter.

3. From understanding social media to new integrative information technologies

If we can understand and model information spreading in e.g. Twitter good enough, then we can measure how newsworthy the content of a tweet is, which, besides the obvious application to journalism, also can help recommender systems. If we can weigh retweets by their importance, then we can also weight co-occurring concepts in the tweet. Co-occurrence is a possible basis for recommender systems—if a user mentions two products in a tweet, they should be more strongly connected (so that if someone likes one product, we can recommend the other). Social media can also be used for monitoring other phenomena in society, and even if Twitter does not spread memes in the same way that infections spread in human communities (like we argue above), it does provide real-time data that can be used to track, for example, influenza outbreaks [17,18]. There are however even simpler digital traces that can predict outbreaks of influenza. In Google Flu Trends [19] Google web searches—which, from the user’s perspective are simply tools to reach a desired web page—are used as data for “predicting the present”, as Google representatives have put it [20]. Thus, information that could be thought to be completely trivial (such as what a randomly selected person has entered into their search engine window) may actually carry considerable predictive value when aggregated on a large scale. Indeed, a recent paper [21] argues that web-search based prediction of things like movie and video game revenues and flu outbreaks perform almost as well as specialized, more or less hand-crafted predictors in general, and in fact better than those predictors when only little information is available. Generalized prediction of consumer behavior (and other types of human behavior) is of course far from a solved problem, as is

the problem of modeling consumer tastes and recommending new products based on those models. Such models can work well on average but fail completely when faced with unusual items. For example, anecdotally, the best-performing algorithms in the NetFlix Prize challenge (<http://www.netflixprize.com>) had considerable trouble classifying offbeat titles like *Napoleon Dynamite* based on the ratings a user had awarded other movies. Also anecdotally, the online platform hunch.com (<http://www.hunch.com>), which builds a kind of personality profiles of users and uses those to offer recommendations on practically any subject, claimed to have predicted with about 95% accuracy how a given person would like *Napoleon Dynamite*. Thus, aggregation of information apparently unrelated to movie ratings may be informative in movie rating prediction. A more well-defined and principled approach to the accurate recommendation of diverse (niche) items was introduced by Zhou et al. [22].

4. Collaborative information technologies: Examples from medicine

We will move from the more theoretical perspective on the science needed for integrating information technologies to look at some examples of practical applications. In particular, we will discuss some collaborative medical technologies where the experience of patients is analyzed by algorithms similar to those discussed above. Revisiting the themes introduced above—the growing importance of data, the new individual-based data economy and the feedback loops between data and behavior—we exemplify by looking at the emerging data-driven, personalized health sector [23]. The nascent field of personal health (or disease) management brings together several recent trends like self-monitoring (measuring data about oneself manually or automatically), online social networks, collaborative filtering and predictive analytics. Collaborative medical sites like PatientsLikeMe (<http://patientslikeme.com>) and CureTogether (<http://curetogether.com>) allow patients to track

and share statistics about their physiological status, medications regimes, mood and so on. By aggregating information over a large number of patients—essentially a large, loosely organized, semi-controlled clinical study—it is possible to discover interesting correlations between symptoms and medications, between different symptoms, or between patient types and symptoms. In fact this is not just a possibility—it has already happened, and such correlations have been validated and described in books, each focusing on a specific disease or symptom, like depression or back pain, put out by CureTogether (<http://curetogether.com/blog/category/books/>). PatientsLikeMe have published several peer-reviewed paper about their platform and findings (e.g. Ref. [24]). From the patient's point of view, the highly quantitative data combined with the social-network design of these sites makes it possible to find other patients with similar symptoms, habits or life situations and to find out which drug and exercise regimens that worked best for them (similar to protein prediction by similarity matching [25]). This adds a social predictive aspect that grows more effective with the increasing size of the patient network. Fig. 4 sketches how data could flow via such a system and feedback into a person's everyday life.

With these new tools, health management becomes simultaneously more community-based and more individually tailored. Instead of working with averages of faceless clinical populations, one gets fine-grained information from which it may be possible to mine more subtle relationships. A bare-bones version of a personal health management site is MyMigraineJournal, where users (migraine sufferers) can try to establish statistically significant relationships between certain pre-defined triggers and migraine episodes. The patients track the occurrence of their trigger events and their migraines, and the site calculates association measures between triggers and migraine using hierarchical Bayesian models. Although this site is

mostly individual-centered, information from the whole population does enter into each patient's own models as Bayesian priors in the model fitting.

5. Personalized medicine

So-called direct-to-consumer genomics companies such as 23andme (<http://www.23andme.com>), Navigenics (<http://www.navigenics.com>) and Knome (<http://www.knome.com>) have also emerged in recent years. They analyze tissue samples submitted by customers and report (among other things) small genetic variations (single-nucleotide polymorphisms, SNPs) on a whole-genome scale. This information, not being very useful in its raw form, can be used as a basis for personal or small-scale research addressing potential effects of an SNP. A newly established company called Genomera (<http://www.genomera.com>; see also Ref. [26]) is launching an online platform meant for the use of small, self-assembled research groups who want to test hypotheses relating to their genotypes. As an example, the article cited above described a group of volunteers who experimented on themselves by trying different medications or combinations of medications during different periods of time, recorded measurements of relevant physiological biomarkers and tried to relate their physiological responses to their SNP genotypes. Although the group was too small for any reliable conclusions to be drawn, this kind of study, if scaled up, could provide useful information.

On the academic and commercial research side, collaborative or individual-based projects are emerging as well. A recent paper in Nature Biotechnology [27] argued in favor of *patient-based* cancer drug discovery complementing the traditional approaches of compound-based or target-based drug discovery. Similar research programs are being implemented by the company CollabRx (<http://www.collabrx.com>), which aims to offer completely individual-based cancer research efforts for customers, including molecular tumor profiling,

bioinformatics and drug optimization. Inspired by the (often ironically used) term for this type of individual-based studies, “N=1 research”, the company N-of-one (<http://www.n-of-one.com/>) provides “a highly customized approach for clients who are battling an active cancer”. The Pink Army Cooperative (<http://pinkarmy.org>) works with individually adapted drug development for breast cancer patients based on synthetic biology. It should be said that these efforts are in very early stages and that they may not turn out to work—but the fact that they exist is interesting in itself.

Further developments in this space could include personalized physiology simulations. Already,

Fig. 1. Retweet distribution from Kwak et al. (2010). The d is the depth of the retweet tree, s are the number of retweeters in a retweet tree. The values are cumulative distributions, i.e. an ordinate value shows the probability of finding a d or s value larger than the value of the abscissa.

Fig. 2. Hypothetical mechanism for limiting the sizes of retweet trees. The shades of the nodes represent their interests (in a one-dimensional space). The network is assortative with respect to the interests (i.e. there is a tendency for nodes of similar shades to connect to each other). Assume the user in the bottom left corner post tweet reflecting his or her interests,

some companies (like Entelos [<http://www.entelos.com>] and Optimata [<http://www.optimata.com>]) offer virtual patient simulators for optimizing clinical trials, and an ambitious European Union project, Virtual Physiological Human (<http://www.vph-noe.eu>), is working towards whole-body simulations. These existing tools could be further augmented by including actual patient data obtained through self-monitoring. In this way, patients could dynamically observe e.g. atherosclerotic build-up in arteries [23].

then other people are more likely to retweet the post the more similar their interests are to the original post. The width of the lines in panel B represents the probability of the original tweet to be transmitted along that edge. The circle segments sketches the effect on the probability the original tweet propagates to some part of the graph (the values are hypothetical, not the result from some model).

Fig. 3. Output of our sketchy retweet model that takes the mechanism of Fig. 2 into account. Curves are based on 100 network realizations and 1000 infection seeds. Panel A shows the epidemic threshold via the average retweet-event size as a function of the transmissibility (the chance of a disease transmission, for a contact, per time unit). The epidemic threshold is for transmissibility of around 0.05. Panel B shows the probability distribution of retweet-event sizes with a cut-off around 200.

Fig. 4. A sketch of how a data-driven health/disease management system might work. An individual (perhaps a patient) keeps track of relevant everyday events or variables (like disease symptoms, physiological measurements, sleep, or food intake) and submits some of these to a social health data platform. The data are analyzed, aggregated and put into the context of the whole population of users. Users get recommendations not only for what behaviors to modify based on their personal measurements and criteria, but also based on what has worked well for similar users. Finally, the aggregated data can be used for explorative correlation studies or calibrating statistical models of (e.g.) diseases.

Conclusions

In the developed societies, we are living in an era of change, where the visions of an information age from forty years ago are becoming our everyday life. Our daily activities produce information that can be analyzed for the benefit of our decisions of what to do. This feedback loop will, we believe, be an important social force in the future, and to make this force a positive one, we need to understand such processes. Some of these phenomena have been described in terms of imprecise analogies—our examples are spreading processes like retweets and video recommendations that often are described, and sometimes modeled, like disease spreading, diffusion or flow. We believe that we have to let go of analogies in our quest for understanding of the feedback loop between information technology and behavior. Different communication platforms have their particular functionality, which create different dynamics, jargon, behavior and social norms, something observed 20 years ago in the context of Internet Relay Chat “it cannot be completely explained or analysed by reference to the methods used by other [computer-mediated-communication] theorists” [28]. If

one are able to understand and model information spreading in social media one can weight the information properly which would be useful for e.g. recommender systems.

Medicine is one of the fields where most integrating and participatory information technologies have been developed. This is maybe not so surprising as the practice of medical doctors is much based on the collected experience of others. If one can integrate a larger body of medical experiences (both from doctors and directly from patients) and other information sources, it is obvious that one can achieve higher precision in diagnoses. This is not our idea alone, and indeed, there are already several new interesting corporate and academic initiatives to use the datasets of human activity for the benefit of e.g. personalized medicine. These applications are inspired by developments in machine learning and bioinformatics rather than clinical practice and many of them could probably have been realized a decade ago using clinical data sets. So in a way, these applications owes their existence to modern information science—new media and their applications need inspiration from the future, not from the past, both to be understood and developed.

Acknowledgements This research is supported by the Swedish Research Foundation and the WCU program through NRF Korea funded by MEST R31-2008-10029.

References

- [1] Toffler A. *Future shock*. Random House, New York, 2007.
- [2] Palme J. You have 134 unread mail! Do you want to read them now? In *Computer-Based Message Services*, Smith H T (ed.), 1984, 175-188.
- [3] Toffler A. *Powershift: Knowledge, Wealth and Violence at the Edge of the 21st Century*. Bantam Books, New York, 1990.
- [4] Lazer D, et al. *Computational Social Science*. Sci-

ence, 2010, 323: 721-723.

[5] Kaplan A M, Haenlein M. Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 2009, 53(1): 59-68.

[6] Goldbeck J. Generating predictive movie recommendations from trust in social networks. *Lecture Notes in Computer Science*, 2006, 3986: pp. 93-104.

[7] Seth A, Zhang J. A social network based approach to personalized recommendation of participatory media content. *International Conference on Weblogs and Social Media, AAAI*, 2007, pp. 109-117.

[8] Watts D J. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company, New York, 2003.

[9] Grönlund A, Holme P. A network-based threshold model for the spreading of fads in society and markets. *Advances in Complex Systems*, 2005, 8: 261-273.

[10] Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web*. 2010, pp. 591-600.

[11] Newman M E J. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 2005 46: 323-351.

[12] Leynaert B, Downs A M, Vincenzi I. Heterosexual transmission of Human Immunodeficiency Virus: Variability of infectivity throughout the course of infection. *American Journal of Epidemiology*, 1998, 148: 88-96.

[13] Anderson R M, May R M. *Infectious diseases in humans*, Oxford University Press, Oxford UK, 1992.

[14] Leicht E A, Holme P, Newman M E J. Vertex similarity in networks. *Phys. Rev. E*, 2006, 73: 026120.

[15] Liben-Nowell D, Kleinberg J. Tracing information flow on a global scale using Internet

chain-letter data. *Proc. Natl. Acad. Sci. USA*, 2008, 105: 4633-4638.

[16] Liljeros F, Edling C R, Amaral L A N. Sexual networks: implications for the transmission of sexually transmitted infections. *Microbes and Infection*, 2003, 5: 189-196.

[17] Lampos V, De Bie T, Cristianini N. Flu Detector – Tracking Epidemics on Twitter. In: *ECML PKDD 2010*, 2010, Springer, pp. 599-602.

[18] Signorini A, Segre A M, Polgreen P M. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE*, 2011, 6(5): e19467.

[19] Ginsberg J, Mohebbi M H, Patel R S, Brammer L, Smolinski M S, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*, 2009, 457: 1012-1014.

[20] Choi H, Varian H. *Predicting the Present with Google Trends*, 2009, Google Inc.

[21] Goel S, Hofman J M, Lahaie S, Pennock D M, Watts D J. Predicting consumer behavior with Web search. *Proc. Natl. Acad. Sci. USA* 2010, 107: 17486-17490.

[22] Zhou T, Kuscsik Z, Liu J-G, Medo M, Wakeling J R, Zhang Y-C. Solving the apparent diversity-accuracy dilemma of recommender systems, *Proc. Natl. Acad. Sci. USA*, 2010 107: 4511-4515.

[23] Swan M. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. *Int. J. Environ. Res. Public Health*, 2009, 2: 492-525.

[24] Wicks P, et al. Sharing health data for better outcomes on PatientsLikeMe. *J. Medical Internet Res.*, 2010, 12(2): e19.

[25] Holme P, Huss M. Role-similarity based functional prediction in networked systems: Application

to the yeast proteome. *J. R. Soc. Interface*, 2005, 2: 327–333.

[26] Dolgin E. Personalized investigation. *Nature Medicine* 2010, 16: 953–955.

[27] Schreiber S L et al. Towards patient-based cancer therapeutics. *Nature Biotechnology*, 2010, 28: 904–906.

[28] Reid E M. *Electropolis: Communication And Community On Internet Relay Chat*. Honours Thesis, 1991, University of Melbourne.

Petter Holme has a M.A. in Chinese (2000), M.Sc. in Engineering Physics (2000) and a Ph.D. in theoretical physics (2004). Holme was as the first Swedish physicist with a dissertation on networks. After postdoctoral positions in the U.S. (Department of Physics, Michigan University and Department of Computer Science, University of New Mexico) and an assistant professorship at the Royal Institute of Technology is Holme associate professor at Umeå University. He leads a research team of five people and is director of the interdisciplinary research environment IceLab. Holme has about 70 scientific publications. His current research interests concern temporal networks, network epidemiology, social media and disaster informatics.

Mikael Huss received the B.A. degree in Chinese language from Lund University, Sweden, in 1998, the M.Sc. degree in Molecular Biotechnology from Uppsala University, Sweden in 2000, and the Ph.D. degree in Computer Science from the Royal Institute of Technology, Stockholm, Sweden in 2007. He is currently working as a bioinformatics scientist at Science for Life Laboratory, a center for high-throughput biomedical science, in Stockholm, Sweden. He has previously held research positions at Genome Institute of Singapore and is currently also affiliated with Stockholm University, Sweden. His research interests lie in, e.g., analyzing massive data sets from high-throughput DNA sequencing, "big data" in general, and network theory. He is the author or co-author of some twenty research articles.