Data Science 1 (2017) 85–94
DOI 10.3233/DS-170003
IOS Press

Maintaining intellectual diversity in data science

Richard P. Mann [a,*] and Olivia Woolley-Meza [b]
[a] *Department of Statistics, School of Mathematics, University of Leeds, UK*
ORCID: https://orcid.org/0000-0003-0701-1274
[b] *Computational Social Science, ETH Zurich, Zurich, Switzerland*
ORCID: https://orcid.org/0000-0003-4517-2765

Editor: Tobias Kuhn (https://orcid.org/0000-0002-1267-0234)
Solicited reviews: Jodi Schneider (https://orcid.org/0000-0002-5098-5667); Melissa Haendel (https://orcid.org/0000-0001-9114-8737); Bjarke Mønsted (https://orcid.org/0000-0003-3683-312X); Victor de Boer (https://orcid.org/0000-0001-9079-039X)

Received 5 March 2017
Accepted 28 March 2017

Abstract. Data science is a young and rapidly expanding field, but one which has already experienced several waves of temporarily-ubiquitous methodological fashions. In this paper we argue that a diversity of ideas and methodologies is crucial for the long term success of the data science community. Towards the goal of a healthy, diverse ecosystem of different statistical models and approaches, we review how ideas spread in the scientific community and the role of incentives in influencing which research ideas scientists pursue. We conclude with suggestions for how universities, research funders and other actors in the data science community can help to maintain a rich, eclectic statistical environment.

Keywords: Collective intelligence, diversity, contagion networks

1. Introduction

In 2012 the Harvard Business Review declared Data Scientist to be the 'sexiest job of the 21st century' [9]. The last five years have borne out that pronouncement. As illustrated in Fig. 1, global interest in data science has increased exponentially, at least as measured by the number of related searches on Google. There has been a huge increase in the number of universities offering courses in 'data science' or 'data analytics', led by student demand in response to a rapid growth in the number of well-paid 'data scientist' job positions. Doubtless, some of this is a relabeling of previously extant activities; much of what we teach our data analytics students has previously been covered in statistics and machine-learning programs, while some companies advertising data scientist positions would have once advertised for statisticians. Nonetheless, within these fields there has also been a substantial increase in activity. To illustrate, the Annual Conference on Neural Information Processing Systems (NIPS), regarded as the

*Corresponding author. E-mail: r.p.mann@leeds.ac.uk.

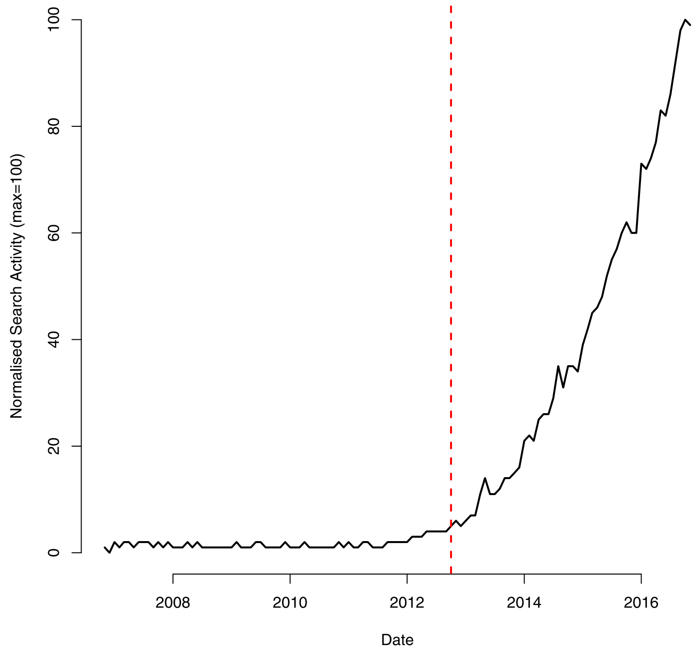

Fig. 1. The monthly search activity for ‘data science’ as a proportion of all Google search activity, extracted from Google Trends between October 2006 and October 2016. Activity is normalized to a maximum of 100 over the time period. The vertical dashed line indicates the date on which The Harvard Business Review declared Data Scientist to be ‘the sexiest job of the 21st century’ [9]. Raw data available in Supplementary Materials.

leading venue for publishing and discussing research in machine-learning, has seen huge year-on-year increases in registrations over the past five years, with nearly 4000 delegates attending in 2015. Overall, it is clear that there has been a step change in the number of individuals and organizations involved in performing sophisticated analyses of large and complex data. And there is no sign yet that this growth in data analytics in industry, government and academia is slowing down. Within scientific research, this huge increase in data analytic capabilities and activity presents us with an important question: how can we ensure that all this work is creating genuine new knowledge? It may seem natural to assume that if more people are trained to analyze data using exciting new tools like Random Forests, Deep Neural Networks and Gaussian Processes, that we should expect more insightful, robust analyses of data to result, and therefore obtain more knowledge from our scientific endeavors. However, in many cases these tools are being applied to data that is statistically troublesome: observational data, often unstructured, subject to strong selection biases, without controls and with many interacting factors potentially affecting the outcome of interest. Examples include text and behavior extracted from social media [37], hospital admission data [25], and store loyalty card records. With rare exceptions [20] these do not permit anything resembling the classic randomized, controlled trials that are the gold standard of causal inference. Moreover, many methods typically employed in machine-learning and industrial data analytics are primarily focused on *predictive* accuracy, rather than inference and interpretation of underlying causal structures. Finally, and importantly for this perspective, data analytic techniques are subject to ‘bubbles’ of interest with the scientific community [42]. In the 1980’s artificial neural networks were firmly at the forefront of machine-learning and artificial intelligence research. The popularity of these waned in 1990’s and 2000’s in favor of Gaussian processes [32], Random Forests [7] and other non-parametric methods, before a resurgence led by new techniques for training many-layered neural networks, termed 'Deep Learning', in the 2010's. These waves of interest in one technique or another have a pronounced effect on the collective scientific enterprise, as they reduce the diversity of statistical models and approaches being used to investigate similar data sets. As such, any problems inherent to a single type of statistical model can be amplified if that technique becomes popular within the community. Meanwhile, the particular advantages of other methods can become lost as their popularity wanes. It is far from clear that the popularity of one method over another is strictly related only to its analytical power; instead there are strong undercurrents of fashion and conformism in the methods researchers are expected to use. Furthermore, conformisms and fashion are further magnified by network effects that give a small number of researchers and methods exponentially more visibility.

Lack of diversity in analytical approach is a missed opportunity. There is substantial cross-disciplinary evidence for the important role diversity plays in collective intelligence [29,40]: experimentally in human [47] and animal behavior [2], in theoretical models of collective behavior [17,24,49], and specifically in the successes of statistical ensemble models (e.g. [4] and see subsequently listed examples). Ensemble models use combinations of multiple, often very many, different statistical models to perform data analysis. For example, the popular Random Forests model [7] is an ensemble model: many distinct decision tree models are generated from a single data set, before being aggregated to make predictions from new data. The advantage of an ensemble approach is that each model within the ensemble may identify and utilize different features and patterns within the data. Aggregated together, the errors made by each model cancel out to some degree, resulting in a collectively accurate prediction. The power of aggregating distinct models for a given data set became evident to many at the conclusion of the Netflix Prize competition [5]. This competition set participants the challenge of improving on the accuracy of Netflix's own algorithm for predicting future film watching choices by at least 10%. For some time no single team working with one statistical model was able to achieve this mark. The competition was eventually concluded when several teams came together, combining their different models so as to improve their overall predictions [6,31,41]. This outcome demonstrated that, in statistical modeling, the search for one 'true' or 'best' model is often misguided. Instead, as a community we should seek the best *combination* of approaches, especially when faced with complex, multi-dimensional phenomena. Since it is rarely possible or desirable to centrally coordinate a search for a good collection of statistical models, we must instead consider how the incentives individuals face and the networks they inhabit influence the type and variety of statistical research that they perform.

1.1. Collective wisdom, collective madness

The ability of a group to exhibit intelligence superior to any of its constituent members is well established. One of the first to study how collective intelligence emerges was Condorcet, who considered the case of an idealized jury [10]. Consider n jurors, tasked with deciding whether a defendant is guilty or innocent. Each juror is individually only accurate in making this determination with a relatively low probability, say 60%. That is little better than guessing at random. However, assuming that the jurors make up their minds independently, Condorcet showed that collective group decision (determined by a simple vote) is far more likely to be accurate than a single individual, growing quickly with the number of jurors, n. Francis Galton made similar observations regarding the accuracy of collective estimation [13], which is ultimately predicated on the Law of Large Numbers: independent errors made by many individuals tend to cancel out in aggregate, making the group much "smarter" than a single individual. As noted above, these early observations have been replicated since within data science, in ensembles of models and research teams.

However, collective wisdom is contingent on diversity and independence between members of a group [40]. The counterpoint to the collective wisdom of Condorcet's idealized jury is the collective madness we see when individuals are too strongly influenced by the collective mood. This is exhibited in famous examples of damaging group think, such as Tulip Fever, the South Sea Bubble and other stock market booms and busts [14]. We may also observe on a daily level in our own lives instances where social conformity and peer pressure lead individuals and groups to behave sub-optimally. The community of researchers in data science, statistics, machine learning and artificial intelligence is also subject to individual, social and insitutional forces that discourage a diversity of approaches [12,23,26,39,48]. Certain statistical models become fashionable, are accepted as the new big thing and are soon ubiquitous. Senior researchers train their students in the methods that they know. At any one time, certain types of model are, generally, more accurate and/or efficient than others, and researchers tend to gravitate towards these. The culture of benchmarking one's new method against the state of the art in terms of accuracy necessitates that researchers utilize the best currently available methodologies if they wish to get their work published. Previous research has shown that when individuals are rewarded solely on the basis of the accuracy of their individual predictions, this tends to lead to a severe lack of diversity and a subsequent catastrophic reduction in collective wisdom, putting the whole collective on a par with a single individual [17,24].

The examples of successful ensemble approaches above demonstrate that diversity has not been extinguished in data science (e.g. [5]). However, in the light of incentives and mechanisms that discourage diversity, and in the knowledge that diversity is critical to the collective wisdom of any community, it behooves us to consider carefully how ideas spread in the research community, and how the incentives and structures inherent to the scientific community can be used to encourage a wider variety of research approaches.

2. Understanding the spread of ideas

Diverse ideas and methods are combined through a decentralized emergent process, as scientists become aware of information, communicate it, "adopt" or use these ideas and create new ones. This process is driven by the dynamic and multi-layered structure of interactions between scientists and the mechanisms via which ideas are taken up. Here we will discuss insights from the study of, complex networks [3,11,28] and spreading processes on these networks [19,30] that can shed light on how to structure scientific interactions in a way that sustains a diverse pool of methods in data science and encourages researchers to combine and integrate them in productive ways.

2.1. *Networks of scientific interaction*

Analyzing the networks of scientific interaction can reveal important information about the current divisions between different methods and the groups of researchers that use them. Furthermore, these network structures also yield a better understanding of the potential for these methods to be better integrated through new interactions. There are many rich sources of data that have recently become available that can be used to characterize the interaction of scientists in different contexts. For example, the Web of Science[1] or other bibliometric sources can be used to construct citation networks. Although this is only an imperfect measure of scientific ideas and how they spread, there is quantitative evidence that scientific

[1] http://clarivate.com/scientific-and-academic-research/research-discovery/web-of-science/

ideas "spread" through citations [21] and that these networks can be used to predict their spread [36]. Using bibliometric data we can also extract the structure of scientific collaborations and co-authorship networks. Similarly we can reveal collaborations and knowledge transfer in other dimensions, for example through the network of user-repository interactions on Github[2] [18]. These networks can be used to understand the potential for spread of ideas and methods between researchers. There are of course other, less formal forms of scientific interaction which are becoming easier to measure. For example, measurements of face-to-face interactions at conferences [38] and exchange and "friendship" on online social platforms such as Twitter and ResearchGate can also be used to map the structure of interactions between scientists.

All of these quantifications of scientific interaction contain different information at different temporal resolutions. However, two pervasive characteristics are community structure and a high inequality, or broad distribution of the centrality (most basically measured through connectedness) of different network components, whether they be researchers or methods.

2.2. Community structure, individual centrality and integrating diverse methods

Modules or communities are, intuitively speaking, the sub units of a network made up of individuals (e.g. researchers or methods) that interact more strongly with each other than with the rest of the network. One interesting approach to integrating diverse perspectives would be to combine methods from distant communities through ensemble methods.

Taking a more decentralized long-term approach, communication channels between researchers operating in these distant communities also need to be established. However, there are many reasons to be wary of a naive approach which simply encourages more unstructured interaction. To retain useful diversity, interaction across communities must not compromise a basic degree of isolation and independence between communities.

The importance of community structure is better understood if we consider the mechanisms that seem to govern the spreading of ideas and innovations. The best studied model is fractional threshold contagion [43], where the probability that an individual becomes "infected" through an infective contact is dependent on the *fraction* of other infected individuals that it interacts with (in network terms, its neighbors). This rule captures the idea that adoption is a social process: there is pressure to conform, or there are added synergistic benefits to adopting if others whom we interact with adopt. Given such a process, communities serve as incubators for new ideas, through local reinforcement. But communities also have the opposite effect, slowing the adoption ideas that originate outside of them. Increasing connectivity randomly throughout a network could decrease the diversity of ideas, allowing only the most contagious, or those that start in the most well connected places, to persist.

Theoretical work indicates that networks that have modular structure but sustain intermediate levels of connectivity between modules provide optimal conditions for the global uptake of ideas [27]. However, such results do not take into account the hierarchical structure of networks, and how the centrality of nodes (for instance most simply according to degree) determines how effectively inter-modular connectivity can enable global spread.

The dynamics of contagion through a fractional threshold mechanisms are driven by the inverse relationship between the connectedness of an individual and its susceptibility to ideas. Thus, more peripheral individuals are more receptive to new ideas and play a key role in sustaining a rich and diverse set of

[2] https://github.com/

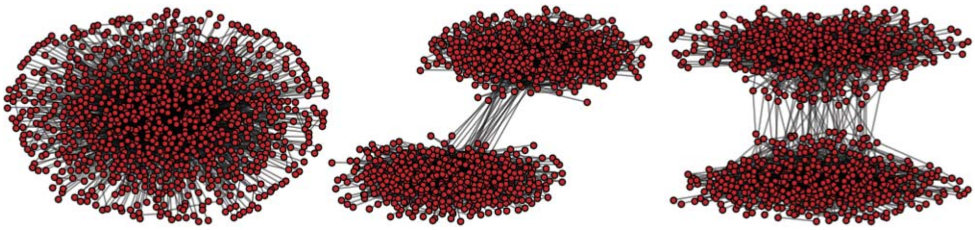

Fig. 2. Three different network topologies that can sustain different diversity of ideas: All networks exhibit stratification, that is, most individuals have a low degree but there is a small number of highly connected individuals. The network on the left has no community structure. This network cannot sustain much diversity. The network in the middle has a strong community structure with inter-community connections through central individuals. In this network the diverse ideas that spark in the different communities cannot spread globally. The network on the right has strong community structure, but with inter-community connections through peripheral individuals. This topology offers a promising approach to sustain the global penetration of diverse ideas that are fostered locally.

scientific ideas. Therefore, in theory ideas can be more effectively transferred from one community to another through connections between more peripheral individuals rather than highly connected individuals (see Fig. 2 for a schematic of the different network topologies we discuss).

To determine how strongly these theoretical conclusions apply to the spreading of scientific ideas and methods a more extensive empirical validation of the prevalence of fractional threshold contagion is needed. Unfortunately, there is a lack of empirical work selecting between plausible information spreading mechanisms in a statistically principled way, in any context. One of the rare examples [34] does in fact offer evidence for the prevalence of fractional threshold contagion – it provides the best fit (from a pool of standard contagion mechanisms) to observed behavioral changes in schooling fish. However, based on studies of information spread on Twitter microblogging [33,45], we would expect that multiple contagion mechanisms operate in the spread of scientific ideas, for instance depending on characteristics of the idea. Interestingly, despite the uncertainty regarding contagion mechanisms, the key role that peripheral researchers can play in the spreading of ideas due to their sensitivity becomes more generally applicable if we look beyond models that only consider one idea spreading in isolation. Ideas are part of an ecosystem, and they interact with each other through their scientific "hosts", akin to the dynamics of co-infection and super-infection studied in evolutionary epidemiology [1]. Competition for limited human time and attention is the most studied interaction [15,16,44]. However, there are also synergistic effects between ideas. Most obviously, ideas that are similar and consistent with each other are more likely to be adopted by the same scientist. These interaction effects accentuate rich-get-richer feedbacks in the system and thus reduce diversity. Furthermore, since the most established scientists tend to be the most connected, they experience more information overload and therefore the competition for their attention is greater, and at the same time, they are less likely to adopt novel ideas that contradict the establishment that has secured their privileged position.

2.3. *The role of incentives*

The type of research that individuals pursue and the methods they use are influenced by their peers and networks, as we have seen above. However, the decisions scientists make in this regard should not be viewed simply as a passive result of community pressure. Instead, these decisions are active choices that are made both to satisfy the researcher's curiosity, but also to achieve their professional objectives such as promotion, funding and recognition. How scientists are rewarded for their research will affect, in positive or negative ways, the diversity of ideas and therefore the collective wisdom of the scientific community.

Scientific ideas and publications exist in a quasi-market, where some are accorded a high value and attract high rewards. A typical researcher is unlikely to pursue highly novel but unpopular ideas if there is no potential for them to be recognized through the awarding of prizes, promotion or further research funding if those ideas later turn out to be fruitful. Likewise, if rewards are systematically allocated to incremental and low-risk research, this will tend to attract more researchers to these areas. Furthermore, researchers are aware of the high premium placed on clear, statistically unambiguous results, and may self-censor weaker findings either by not publishing or by not making data freely available [46]. This is in addition to the external restrictions imposed by journals unwillingness to publish work that does not report a statistically significant positive finding [48].

When individual rewards are oriented solely towards simple criteria of success this tends to suppress diversity and risk-taking [17,24]. Conversely, there is recent evidence that useful diversity can be encouraged by rewarding accurate *minority* predictions [24]. These are occasions on which an individual or a model predicts correctly, while the majority of others predict incorrectly. This creates an incentive to focus on less exploited sources of information, or more niche features of a data set, since the individual cannot win any reward by simply imitating what the majority of others are already doing. For the individual, this may make their model less accurate overall. But it makes their model far *more* useful to the community, as it contributes additional information not already presented by others.

Similar ideas are already applied in established methods of ensemble creation. For example, the technique of boosting [35] is a meta-algorithm for assembling ensembles of weak classifiers that act as a strong classifier in aggregate. A common way to achieve this is to iteratively add weak classifiers to an ensemble, re-weighting the data set under consideration after each new classifier is added. Examples within the data that are currently poorly classified are given higher weights, while those which are already well classified are given lower weights. In this way, additional weak classifiers are 'incentivized' to focus on accurately classifying the examples that are currently poorly modeled.

3. Conclusion

In this paper we have sought to highlight the importance of intellectual diversity in scientific research, especially in relation to data science. We have discussed how the scientific community structure and the individual incentives inherent in research work are likely to affect this diversity. Given the great uncertainty about the mechanisms driving information dynamics in science, we have to be cautious in suggesting what interaction structures can best sustain diversity of methods in data science and their productive and eclectic integration. However, some guiding principles are clear.

Fostering the right kind of networks for collaboration and information transfer is key. Diversity thrives in a network structure that allows communities to work in partial isolation. Work carried out in the borders between communities and at the periphery of the establishment can lead to important innovations. Therefore enough funding and other forms of incentives need to be allocated to these regions. This can take the form, for example, of funding currently unfashionable research, promoting smaller workshops focused on niche areas or supporting early career researchers who create or utilise non-mainstream methodologies, rather than pushing them to follow the herd.

Detecting existing communities of methods and promoting their integration is an opportunity to improve the power of methods. This can be successfully achieved through formally bringing together different models in ensembles, as discussed previously. However, these final combinations are often the end stage of a much less formal process where individuals become aware of the existence and utility of other methods. Beyond planned combination of methods, connections that sustain interchange between separate communities, and from the periphery to the core of the research community are necessary. However, these connections must be well timed, since premature competition with more established ideas can be counterproductive. Furthermore, the individuals in the core, who accumulate connections and prestige, are not necessarily the most effective integrators. These individuals are most embedded in the currently leading methodologies and they have the most to lose when established approaches are overturned. Thus, increasing connections between scientists working at the periphery, in communities that are typically distant, could be a promising new way of fostering a diverse set of ideas and integrating them for innovative science.

All of these ideas concerning scientific networks are intimately connected to the incentive structures of the scientific community. Most obviously, the smaller communities of researchers focused on less fashionable research need support to carry out their work. However, we also must not lose sight of the fact that the allocation of funding and other scientific rewards such as career progression is not simply a mechanism for enabling researchers with varied interests, but also acts as a driver of those interests. Most researchers wish for some degree of recognition and reward for their work, and will gravitate to areas that offer this. By reforming rewards to encourage diversity, for instance by explicitly favoring minority research ideas, we can avoid wasteful and potentially damaging group think and maintain the rich variety of data analytic approaches. In the data science and machine learning communities this may mean setting less importance on how well a new model is able to predict unseen data, and more importance on how differently it does so compared to existing methods.

Much research remains ahead of us. We need better mapping of the structure of scientific interactions by aggregating information from the different sources available. Of special interest is measuring and characterizing communities that capture the time-varying structure of scientific interaction. This will enable a faster and more precise identification of the scientists and methods emerging in new communities and in community boundaries. This in turn would allow us to foster emerging innovations and identify ripe areas of research which would benefit from contact with the mainstream. However, we must remember the limits of predicting scientific innovation [8]. Scientific approaches cannot be evaluated entirely on their immediate or short-term performance. As first postulated by Kuhn [22], paradigm changing ideas are accepted because of their yet unverified potential and because the social dynamics support change. Rather than over-engineering systems for ‘picking winners’ among new ideas, we should instead focus on removing structures that depress diversity, and embrace the power of collective wisdom.

Acknowledgements

O.W.-M. acknowledges that this work was partially funded by the European Community’s H2020 Program under the funding scheme “FETPROACT-1-2014: Global Systems Science (GSS)”, grant agreement 641191 “CIMPLEX: Bringing CItizens, Models and Data together in Participatory, Interactive SociaL EXploratories” (http://www.cimplex-project.eu).

References

[1] S. Alizon, Co-infection and super-infection models in evolutionary epidemiology, *Interface Focus* **3**(6) (2013), 20130031. doi:10.1098/rsfs.2013.0031.

[2] L.M. Aplin, D.R. Farine, R.P. Mann and B.C. Sheldon, Individual-level personality influences social foraging and collective behaviour in wild birds, *Proceedings of the Royal Society B* **281**(1789) (2014), 20141016. doi:10.1098/rspb.2014.1016.
[3] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, *Science* **286**(5439) (1999), 509–512. doi:10.1126/science.286.5439.509.
[4] E. Bauer and R. Kohavi, An empirical comparison of voting classification algorithms: Bagging, boosting, and variants, *Machine Learning* **36**(1–2) (1999), 105–139. doi:10.1023/A:1007515423169.
[5] R.M. Bell and Y. Koren, Lessons from the Netflix prize challenge, *ACM SIGKDD Explorations Newsletter* **9**(2) (2007), 75–79. doi:10.1145/1345448.1345465.
[6] R.M. Bell, Y. Koren and C. Volinsky, The Bellkor solution to the Netflix prize, 2007. http://brettb.net/project/papers/2007%20The%20BellKor%20solution%20to%20the%20Netflix%20prize.pdf.
[7] L. Breiman, Random forests, *Machine Learning* **45**(1) (2001), 5–32. doi:10.1023/A:1010933404324.
[8] A. Clauset, D.B. Larremore and R. Sinatra, Data-driven predictions in the science of science, *Science* **355**(6324) (2017), 477–480. doi:10.1126/science.aal4217.
[9] T.H. Davenport and D. Patil, Data scientist: The sexiest job of the 21st century, *Harvard Business Review* **90**(10) (2012). https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.
[10] N. De Condorcet, *Essai sur L'application de L'analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix*, Cambridge University Press, 2014. ISBN-10:1108077994 / ISBN-13:9781108077996.
[11] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, Cambridge University Press, 2010. ISBN:9780521195331.
[12] A.M. Edwards, R. Isserlin, G.D. Bader, S.V. Frye, T.M. Willson and H.Y. Frank, Too many roads not taken, *Nature* **470**(7333) (2011), 163–165. doi:10.1038/470163a.
[13] F. Galton, Vox populi, *Nature* **75**(7) (1907), 450–451. doi:10.1038/075450a0.
[14] P.M. Garber, Famous first bubbles, *The Journal of Economic Perspectives* **4**(2) (1990), 35–54. doi:10.1257/jep.4.2.35.
[15] J.P. Gleeson, K.P. O'Sullivan, R.A. Baños and Y. Moreno, Effects of network structure, competition and memory time on social spreading phenomena, *Physical Review X* **6**(2) (2016), 021019. doi:10.1103/PhysRevX.6.021019.
[16] N.O. Hodas and K. Lerman, How visibility and divided attention constrain social contagion, in: *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, IEEE, 2012, pp. 249–257. doi:10.1109/SocialCom-PASSAT.2012.129.
[17] L. Hong, S.E. Page and M. Riolo, Incentives, information, and emergent collective accuracy, *Managerial and Decision Economics* **33**(5–6) (2012), 323–334. doi:10.1002/mde.2560.
[18] Y. Hu, J. Zhang, X. Bai, S. Yu and Z. Yang, Influence analysis of Github repositories, *SpringerPlus* **5**(1) (2016), 1268. doi:10.1186/s40064-016-2897-7.
[19] M.J. Keeling and K.T. Eames, Networks and epidemic models, *Journal of the Royal Society Interface* **2**(4) (2005), 295–307. doi:10.1098/rsif.2005.0051.
[20] A.D. Kramer, J.E. Guillory and J.T. Hancock, Experimental evidence of massive-scale emotional contagion through social networks, *Proceedings of the National Academy of Sciences* **111**(24) (2014), 8788–8790. doi:10.1073/pnas.1320040111.
[21] T. Kuhn, M. Perc and D. Helbing, Inheritance patterns in citation networks reveal scientific memes, *Physical Review X* **4**(4) (2014), 041036. doi:10.1103/PhysRevX.4.041036.
[22] T.S. Kuhn and D. Hawkins, The structure of scientific revolutions, *American Journal of Physics* **31**(7) (1963), 554–555. doi:10.1119/1.1969660.
[23] J. Lorenz, H. Rauhut, F. Schweitzer and D. Helbing, How social influence can undermine the wisdom of crowd effect, *Proceedings of the National Academy of Sciences* **108**(22) (2011), 9020–9025. doi:10.1073/pnas.1008636108.
[24] R.P. Mann and D. Helbing, Optimal incentives for collective intelligence, *Proceedings of the National Academy of Sciences* (2017). doi:10.1073/pnas.1618722114.
[25] R.P. Mann, F. Mushtaq, A.D. White, G. Mata-Cervantes, T. Pike, D. Coker, S. Murdoch, T. Hiles, C. Smith, D. Berridge, S. Hinchliffe, G. Hall, S. Smye, R.M. Wilkie, J.P.A. Lodge and M. Mon-Williams, The problem with big data: Operating on smaller datasets to bridge the implementation gap, *Frontiers in Public Health* **4** (2016), 248. doi:10.3389/fpubh.2016.00248.
[26] M. Moussaïd, Opinion formation and the collective dynamics of risk perception, *PLoS ONE* **8**(12) (2013), e84592. doi:10.1371/journal.pone.0084592.
[27] A. Nematzadeh, E. Ferrara, A. Flammini and Y.-Y. Ahn, Optimal network modularity for information diffusion, *Physical Review Letters* **113**(8) (2014), 088701. doi:10.1103/PhysRevLett.113.088701.
[28] M. Newman, *Networks: An Introduction*, OUP, Oxford, 2009. ISBN:9780199206650.
[29] S.E. Page, *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*, Princeton University Press, 2008. ISBN:9780691138541.
[30] R. Pastor-Satorras, C. Castellano, P. Van Mieghem and A. Vespignani, Epidemic processes in complex networks, *Reviews of Modern Physics* **87**(3) (2015), 925. doi:10.1103/RevModPhys.87.925.

[31] M. Piotte and M. Chabbert, The Pragmatic Theory solution to the Netflix grand prize. Netflix prize documentation, 2009. http://www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf.
[32] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006. ISBN-10:026218253X / ISBN-13:978-0262182539.
[33] D.M. Romero, B. Meeder and J. Kleinberg, Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter, in: *Proceedings of the 20th International Conference on World Wide Web*, ACM, 2011, pp. 695–704. doi:10.1145/1963405.1963503.
[34] S.B. Rosenthal, C.R. Twomey, A.T. Hartnett, H.S. Wu and I.D. Couzin, Revealing the hidden networks of interaction in mobile animal groups allows prediction of complex behavioral contagion, *Proceedings of the National Academy of Sciences* **112**(15) (2015), 4690–4695. doi:10.1073/pnas.1420068112.
[35] R.E. Schapire, The boosting approach to machine learning: An overview, in: *Nonlinear Estimation and Classification*, Springer, 2003, pp. 149–171. doi:10.1007/978-0-387-21579-2_9.
[36] F. Shi, J.G. Foster and J.A. Evans, Weaving the fabric of science: Dynamic network models of science's unfolding structure, *Social Networks* **43** (2015), 73–85. doi:10.1016/j.socnet.2015.02.006.
[37] V. Spaiser, T. Chadefaux, K. Donnay, F. Russmann and D. Helbing, Communication power struggles on social media: A case study of the 2011–12 Russian protests, *Journal of Information Technology & Politics* (2017). doi:10.1080/19331681.2017.1308288.
[38] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J.-F. Pinton, N. Khanafer, W. Van den Broeck et al., Simulation of an seir infectious disease model on the dynamic contact network of conference attendees, *BMC Medicine* **9**(1) (2011), 87. doi:10.1186/1741-7015-9-87.
[39] P.E. Stephan, Research efficiency: Perverse incentives, *Nature* **484**(7392) (2012), 29–31. doi:10.1038/484029a.
[40] J. Surowiecki, *The Wisdom of Crowds*, Random House LLC, 2005. ISBN-10:0385721706 / ISBN-13:978-0385721707.
[41] A. Töscher, M. Jahrer and R.M. Bell, The Bigchaos solution to the Netflix grand prize. Netflix prize documentation, 2009. http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf.
[42] X. Wang and A. McCallum, Topics over time: A non-Markov continuous-time model of topical trends, in: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2006, pp. 424–433. doi:10.1145/1150402.1150450.
[43] D.J. Watts, A simple model of global cascades on random networks, *Proceedings of the National Academy of Sciences* **99**(9) (2002), 5766–5771. doi:10.1073/pnas.082090499.
[44] L. Weng, A. Flammini, A. Vespignani and F. Menczer, Competition among memes in a world with limited attention, *Scientific Reports* **2**, 335 (2012). doi:10.1038/srep00335.
[45] L. Weng, F. Menczer and Y.-Y. Ahn, Virality prediction and community structure in social networks, *Scientific Reports* **3**, 2522 (2013). doi:10.1038/srep02522.
[46] J.M. Wicherts, M. Bakker and D. Molenaar, Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results, *PloS One* **6**(11) (2011), e26828. doi:10.1371/journal.pone.0026828.
[47] A.W. Woolley, C.F. Chabris, A. Pentland, N. Hashmi and T.W. Malone, Evidence for a collective intelligence factor in the performance of human groups, *Science* **330**(6004) (2010), 686–688. doi:10.1126/science.1193147.
[48] N.S. Young, J.P. Ioannidis and O. Al-Ubaydli, Why current publication practices may distort science, *PLoS Medicine* **5**(10) (2008), e201. doi:10.1371/journal.pmed.0050201.
[49] A. Zafeiris and T. Vicsek, Group performance is maximized by hierarchical competence distribution, *Nature Communications* **4** (2013). doi:10.1038/ncomms3484.