International Journal of Internet Science
2012, 7 (1), 1–5 ISSN 1662-5544

International Journal of Internet Science

IJIS.NET

"Big Data": Big Gaps of Knowledge in the Field of Internet Science

Chris Snijders[1], Uwe Matzat[1], Ulf-Dietrich Reips[2,3]

[1]Eindhoven University of Technology, The Netherlands,
[2]University of Deusto, Spain, [3]IKERBASQUE, Basque Foundation for Science, Spain

---

As a member of the editorial board and editors of the International Journal of Internet Science we would like to take this opportunity to comment on some interesting developments in the field of Internet/Web Science. The analysis of so-called "Big Data" has received a remarkable momentum. Conducting a search with "Big Data" as a query, we find 130 entries in the ISI Web of Science, as of July 12. Of these, 94 publications have appeared since 2008, which is no surprise because before 2008 there was no terminological consensus. For 2008 through 2011 the number of publications in the ISI Web of Science equals 16, 16, 13, and 26. In the first half of 2012 (July 12), we find 23 publications for just the first six months, suggesting a rapid future increase. As we explain below, this stream of research has provided useful insights, but also suffers from some serious limitations. The interesting point is that these limitations can (and have to) be addressed by theory guided research that is typically conducted by social scientists. Accordingly, opportunities emerge for those social and behavioral scientists who are willing to collaborate with the Big Data researchers in the natural, engineering, and computer sciences. While this short editorial does not claim to provide an exhaustive overview of Big Data research we hope that it contributes to clarifying what type of questions and problems need input from the social and behavioral sciences. We have the feeling that these knowledge gaps have not yet received the attention that they deserve, and would certainly welcome submissions along these lines in this journal.

What we know

Big Data is a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard statistical software. The rise of digital and mobile communication has made the world become more connected, networked, and traceable and has typically lead to the availability of such large scale data sets (Rainie & Wellman, 2012). Some of the keepers of Big Data sets develop interfaces for everyone to access and analyze some of the data, e.g. Google provides freely available Google Insights, while others hesitate to offer any access. Scientists have begun to develop Web services with interfaces to collectors of Big Data sets, e.g., Milne and Witten (2009) for Wikipedia at http://wikipedia-miner.cms.waikato.ac.nz/ and Reips and Garaizar (2011) for Twitter at http://tweetminer.eu.

In this editorial, we focus on the stream of Big Data analysis that considers different kinds of online (and offline) networks. Analyses of different kinds of networks have shown that in many empirical networks the distribution of the degrees of the nodes follows a power-law (e.g., Barabasi, Albert, & Jeong, 2000). Another network characteristic that has received a lot of attention is the extent to which a network can be considered "small

---

Address correspondence to Chris Snijders, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands, c.c.p.snijders@tue.nl and the International Journal of Internet Science, ijis@ijis.net
The present work was supported by COST Action 1004 "WEBDATANET" (http://webdatanet.eu)

world": pairs of nodes have a low shortest path length between them and the network as a whole is typically organized as a set of dense but loosely connected clusters (Watts & Strogatz, 1998). More recent findings include dynamic properties of networks such as whether networks have a constant average degree (the number of edges growing linearly with the number of nodes) or whether the diameter of the network decreases over time. The empirical networks under study range from the World Wide Web (Barabasi, Albert, & Jeong, 2000), science citation networks (Leskovec, Kleinberg, & Faloutsos, 2007), sexual relationships (Liljeros et al., 2001), to telephone networks (Cortes & Pregibon, 2001).

Given such often observed empirical regularities, several micro-models have been suggested that might lead to networks with the desired properties as suggested above. Some well known models are the Erdos-Rényi random graph model (Erdos & Rényi, 1959), the small-world model (Watts & Strogatz, 1998), preferential attachment (Price, 1976; Yule, 1925), the edge copying model (Kleinberg et al., 1999), and community guided attachment and forest fire models (Leskovic, Kleinberg, & Faloutsos, 2007). Research in this area shows, for instance, that when we assume that each new node connects to existing nodes with a probability that is proportional to the degree of the existing node – the key assumption in the preferential attachment model – that one then indeed ends up with networks that have degree distributions that follow a power law. The underlying logic is compelling: if we find that many real world networks have property X, let us try to understand which processes could lead to property X.

What we do not know: social science theories as guidance for the analysis of micro-processes leading to macro-outcomes

A crucial problem is that we do not know much about the underlying empirical micro-processes that lead to the emergence of these typical network characteristics of Big Data. Most of the underlying process models at the node level are inspired by mathematical ease of exposition, tractability or quite crude approximations of what could really be going on. For instance, the basic preferential attachment model assumes that existing nodes do not connect to each other at all (no new ties between those already in the network). This, however, is a strong assumption that has never been tested adequately to find out whether its violation leads to divergent outcomes at the macro-level of the whole network. Instead of trying to find micro-processes that lead to certain aggregate network properties based on mathematical tractability, one could follow a different analytical strategy and try to come up with micro-processes that match with actual behavior. And this is exactly where social and behavioral research can play its role.

To gain knowledge about the underlying micro-processes social scientists could consider several (online) social networks and measure the process of tie-formation in more detail, derive network micro-foundations from these measurements, and then consider the network properties that follow from it. It is unclear whether this is possible for all types of online networks, but there are certainly more than enough opportunities to consider. For instance, the micro-processes of blog networks and the micro-processes of posting behavior within knowledge sharing online communities (such as emailing lists) lend themselves to such an approach. Both types of networks have been objects of mathematical modeling (Cointet & Roth, 2009; Goetz et al, 2009). However, it is unclear whether, and if so, to what extent the models' assumptions rest on realistic mechanisms that take place at the micro level during the tie-formation, and the suggested approach would complement the mathematical method perfectly. The crucial addition to the literature rests in the fact that such an approach utilizes not only the data on nodes and their interconnections. In addition, survey and interview data about characteristics of the actors and the characteristics of the online community as a whole can be collected and combined with the online data.

As the starting point for such an endeavor, one could consider empirical sociological and social-psychological analyses of processes of tie-formation and bring these back to a limited number of behavioral mechanisms, such as homophily of different kinds, reciprocity, scope of access to other nodes, etc. This knowledge can then be used as input for the selection and formulation of mathematically tractable models of tie-formation. Such an approach would create that empirical analyses rest on sociological and social-psychological theories of network evolution that have been argued and validated empirically instead of on micro-models that happen to be tractable. For sure, such theories are readily available in the social sciences: it has been argued that tie-formation is guided by considerations of reciprocity (Schnegg, 2006), or by homophily with respect to gender or status (Shrum, Cheek, & Hunter, 1988; Hipp & Perrin, 2009), by balance and transitivity ('a friend of a friend is my friend', see Kossinetts and Watts, 2009), or by complex propagation (Centola, Eguiluz, & Macy, 2007). In addition, actor characteristics, such as visibility in the network or in general displaying an attractive trait (such as high reputation or high status) could determine (incoming) tie-formation (Snijders & Weesie, 2009; Stephen & Toubia, 2009), refining the preferential attachment argument.

In short, the crucial point is that the combination of large but sparse Big Data with smaller but rich survey data offers the opportunity to link the individual-level and the community-level characteristics with the individual online data. In addition, one can then study whether the micro-mechanisms that steer tie-formation differ for online communities of different size, coherence, etc. The results of such analyses could then inform the selection of mathematical models about micro-processes.

Why we should know

Understanding networks and network formation is a core topic in complexity research and its underlying sociological and social-psychological processes should receive more attention in the analysis of Big Data for a number of reasons. Some online networks have typical characteristics that are desirable for specific practical purposes. For instance, small world networks are regarded as robust against random damages, but they are vulnerable to coordinated, intentional attacks. Random networks, on the other hand, hardly provide targets for coordinated attacks that could lead to extreme damages in characteristics such as average path length (Barabasi, 2003). Knowledge about the micro-mechanisms and conditions for the emergence of online network characteristics would be an important input for coordinated efforts to stimulate the emergence of these desired characteristics. Furthermore, many argue that the combination of Big Data efforts with social science theory would be useful for the prediction of social and economic crises (Helbing & Balietti, 2011; Conte, Gilbert, Bonelli, & Helbing, 2011). The FuturICT project (Bishop, Helbing, Lukowicz, & Conte, 2011) is an outcome of (and a starting point for) researchers in several countries who share these hopes. The editors of the International Journal of Internet Science, of course, follow these endeavours with great interest and welcome high quality submissions that tackle issues such as the ones mentioned here.

The current issue

Issue 1 of Volume 7 includes four peer-reviewed research articles and one (non-peer-reviewed) supplement of the WEBDATANET research network (EU COST Action 1004). Nic Newman (Reuters Institute for the Study of Journalism), William H. Dutton and Grant Blank (both Oxford Internet Institute) open the stage with their article *Social Media in the Changing Ecology of News: The Fourth and Fifth Estates in Britain* that makes no less than the claim that participation via the Internet has become a Fifth Estate complementing legislature, judiciary and executive. The Fifth Estate is seen as undermining the power of the press (the Fourth Estate). Newman et al. empirically support their analysis by combining multiple methods, including surveys, log file analyses, and interviews.

In the second article, *Putting the "Fun Factor" into Gaming: The Influence of Social Contexts on Experiences of Playing Videogames*, Linda Kaye (Edge Hill University, Lancashire) and Jo Bryce (University of Central Lancashire, Preston) examine social processes in online gaming. Making use of focus group data, the authors present findings that point to the importance of social belonging and social networking for game enjoyment. Moreover, their results suggest that in addition to the well-known phenomenon of (individual) flow a collective experience of flow which they call group flow, may occur in social gaming contexts.

Then, Christopher R. Wolfe, Christopher R. Fisher (both Miami University), Valerie F. Reyna (Cornell University) and Xiangen Hu (The University of Memphis) present three Web experiments in their article, *Improving Internal Consistency in Conditional Probability Estimation with an Intelligent Tutoring System and Web-Based Tutorials*. Fuzzy-trace theory was shown to correctly predict performance in all three experiments for sets of problems involving probabilities. All Web-based tutorials (based on 2x2 tables, Euler diagrams and AutoTutor, a Web-based system with talking animated agents that converse with learners using Latent Semantic Analysis to "understand" natural language) were successful, with different strengths for different types of problems. The results of the Web experiments provide important insights on how Bayesian reasoning can be improved using appropriate tutorials. The research has potentially wide-reaching applications in education.

In the fourth article, *Sharing Only Parts of Me: Selective Categorical Self-Disclosure Across Internet Arenas*, Alison Attrill (De Montfort University, Leicester) investigates the reported content of online disclosures in four different Internet arenas, social networking, instant messaging, general communication, and online shopping. Using a self-disclosure scale amongst a sample of students to measure the revelation of information pertaining to individuals' beliefs, relationships, personal matters, interests, and intimate feelings, the current findings show that self-disclosure on the Internet is more categorical and goal-directed than can be accounted for by existing theoretical explanations of online self-disclosure.

Finally, members of the European WEBDATANET research network on Web-based data collection present their project and kindly invite researchers from the International Journal of Internet Science's readership to participate.

Acknowledgements and Impact (now >3.625, ISI: >2.625)

The editors of the International Journal of Internet Science would like to thank the many individuals and institutions that provided help and support. First of all, and almost seven years by now, we are very thankful for Dr. Frederik Funke's support and help. Without his devotion, time, and constructive critical comments the journal would not have reached its high quality. We also thank Andrés Jiménez from iScience group at University of Deusto, in particular for setting up the journal's new Facebook page at http://www.facebook.com/pages/International-Journal-of-Internet-Science/251579034934885 (please "like" us). We acknowledge the institutional support of the University of Deusto, the Eindhoven University of Technology, the ZPID Leibniz-Institute for Psychology Information and the GESIS – Leibniz-Institute for the Social Sciences. Moreover, we are grateful to the German Society for Online Research (Deutsche Gesellschaft für Online Forschung, DGOF) for its financial support and especially to Lars Kaczmirek (member of the DGOF board) for his collaboration during the GOR 12 conference.

The International Journal of Internet Science was recently added to the list of journals in the Academia.edu network at http://journals.academia.edu/InternationalJournalOfInternetScience. It is currently under evaluation by Thomson Reuters for inclusion with its ISI Web of Science database. As an indicator for possible upcoming inclusion the ISI Web of Science began listing the acronym INT J INTERNET SCI in citations to the journal. According to Google Scholar Citations, where the International Journal of Internet Science has a page at http://scholar.google.com/citations?user=OCYy1o4AAAAJ, citations have increased substantially (because more citing articles published in 2011 to IJIS articles that appeared in 2009 and 2010 are now known and can thus be included in the analysis). Based on these figures, the estimate for the 2011 journal minimum impact factor following calculations in the previous editorial (Reips, 2011) is now at **3.625**, and will be **2.625** for publications within the ISI database upon inclusion with ISI Web of Science.

References

Barabasi, A. (2003). *Linked: How everything is connected to everything else and what it means*. New York: Plume.

Barabasi, A., Albert, R., & Jeong H. (2000) Scale-free characteristics of random networks: the topology of the world-wide web. *Physics A Statistical Mechanics and its applications, v281, 1–4*, 69–77.

Bishop, S., Helbing, D., Lukowicz, P., & Conte, R. (2011). FuturICT: FET flagship pilot project. *Procedia Computer Science, 7,* 34–38.

Centola, D, V.M. Eguíluz, & M.W. Macy. 2007. Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications, 374*, 449–456.

Cointet, J.P., & Roth, C. (2009). Socio-semantic dynamics in a blog network. International Conference on Computational Science and Engineering. doi:10.1109/CSE.2009.105

Conte, R., Gilbert, N., Bonelli, G., & Helbing, D. (2011). FuturICT and social sciences: Big Data, big thinking. *Zeitschrift fur Soziologie, 40,* 412–413.

Cortes, C. & Pregibon, D. (2001) Signature-based methods for data streams. *Journal of Knowledge Discovery and Data Mining, 5*, 167–182.

Erdős, P. & Rényi, A. (1959). On random graphs. I. *Publicationes Mathematicae, 6*, 290–297.

Goetz, M., J. Leskovec, M. McGlohon, & C. Faloutsos (2009) Modeling blog dynamics. *AAAI Conference on Weblogs and Social Media*, 2009. Retrieved from http://cs.stanford.edu/~jure/pubs/blogs-icwsm09.pdf

Helbing, D. & Balietti, S. (2011). From social data mining to forecasting socio-economic crises. *European Physical Journal-Special Topics, 195,* 3–68.

Hipp, J. R. & Perrin, A. J. (2009). The simultaneous effect of social distance and physical distance on the formation of neighborhood ties. *City & Community, 8,* 5–25.

Kleinberg, J.M., R. Kumar, P. Raghavan, S. Rajagopalan, & A. S. Tomkins. (1999) The Web as a graph: Measurements, models and methods. *Lecture Notes in Computer Science, 1627*, 1–17.

Kossinets, G. & Watts, D.J. (2009). Origins of homophily in an evolving social network. *American Journal of Sociology, 115*, 405–450.

Leskovic, J., J. Kleinberg, & C. Faloutsos (2007) Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1, Article 2.

Liljeros, F., Edling, C.R., Nunes Amaral, L.A., Stanley, H.E., & Aberg, Y. (2001) The web of human sexual contacts. *Nature, 411*, 908–909.

Milne, D., & Witten, I. H. (2009). An open-source toolkit for mining Wikipedia. Procedures of the New Zealand Computer Science Research Student Conference, 9.

Price, D. J. de S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27,* 292–306. doi:10.1002/asi.4630270505

Raine, L., & Wellman, B. (2012). *Networked. The new social operating system*. Cambridge: MIT Press.

Reips, U.-D. (2011). Journal impact revisited. *International Journal of Internet Science, 6*(1), 1–7.

Reips, U.-D., & Garaizar, P. (2011). Mining Twitter: Microblogging as a source for psychological wisdom of the crowds. *Behavior Research Methods, 43*, 635–642. doi:10.3758/s13428-011-0116-6

Schnegg, M. (2006). Reciprocity and the emergence of power laws in social networks. *International Journal of Modern Physics C, 17,* 1067–1076.

Shrum, W., Cheek, N.H., & Hunter, S.M. (1988). Friendship in school: gender and racial homophily. *Sociology of Education*, *61*, 27–39.

Snijders, C. & Weesie, J. (2009). Reputation in an online programmers' market. In K. S. Cook, C. Snijders, V. Buskens, & C. Cheshire (Eds.), *Trust and reputation* (pp. 166–185). New York: Russel Sage Foundation.

Stephen, A. T. & Toubia, O. (2009). Explaining the power-law degree distribution in a social commerce network. *Social Networks, 31,* 262–270.

Watts, D. J.; & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440–442. doi:10.1038/30918

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London*, Ser. B 213: 21–87. doi:10.1098/rstb.1925.0002