The Interplay of Theory and Observation: A Proposition for Structured Research on Human Behavior on the Web

Pascal Jürgens
pascal.juergens@gmail.com
Department of Communication, Johannes Gutenberg University Mainz, Germany

Andreas Jungherr
andreas.jungherr@gmail.com
Department of Political Science, Johannes Gutenberg University Mainz, Germany

Benjamin Heitmann
benjamin.heitmann@deri.org
Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland

Motivation

The attempt of Web Science to develop a deeper understanding of human behavior on and with the web, as practiced today, struggles to transcend the stage of isolated case studies of individual phenomena with little or no connection to the nature of human behavior as a whole. The authors believe this state can be remedied by a more conscious combination of theoretical concepts of human behavior and empirical work. To this end this paper identifies four key challenges in sound Web Science: A - Providing theoretical context for studies, B - addressing the role of technological design and communication culture, C - dealing with large data sets and D - charting the web so research can be placed within. We then propose a blueprint for research practices which is based on the school of critical rationalism and serves to increase a study's contribution to the field of web science.

Beyond examples

Since it is not possible to empirically measure every aspect of any given field, science needs clear theories on the nature of its respective field of interest. Individual scientific studies serve as an evaluation of hypotheses which have been postulated on the basis of these theories. The value of a scientific study therefore lies not in the mere accurate description of its topic but in the connection the scientist establishes between the study and theories of the field.[1] This is the critical rationalists understanding of scientific methodology.[2] A mere positivistic approach to science - that is, the collection of separate valid descriptions of isolated phenomena - does not lead to a deeper understanding of the world.[3]

For web science this means that studies about human behavior on-line have to be put in relation to more general concepts of human behavior. Also the population in question has to be put in relation to the population as a whole. Only this process enables us to gain a deeper understanding of the impact of the web on human behavior and thus answer the basic question of web science. This process has to be based on explicit theoretical concepts about human behavior and the corroboration or refutation of domain specific hypotheses. But before we can incorporate a study into the context of established theories, we have to be able to judge it on its validity. Since web science is an emerging scientific field, with practitioners who come from different disciplines and thus do not necessarily share the same research practices, it is important to address at least the most obvious research challenges.

A staged selection model of validity in web science

Like any field of empirical enquiry, web science has to address liabilities that arise in every endeavor that involves data acquisition and data analysis: *(i) sampling*, *(ii) observational error* and *(iii) inferential error*. With web science, these liabilities appear at several levels and in different combinations, giving rise to four key challenges that we will address later in this paper. All three factors determine the accuracy - and ultimately the validity - of scientific descriptions of human behavior.

(i) Sampling

Sampling is the process that addresses the question "How can we ensure that the results from our survey actually speak for all the people we are talking/ thinking about, not just the ones we measured?"

On the web, sampling is imperative for all studies where the complete data set (such as a log of all the interactions between all the users) is unavailable. Sampling is also implicitly applied if inferences about the population of the web (p_w) or even the general population (p_g) are made. If a sample is not selected appropriately, it may carry a sample bias and thus prohibit valid inferences about the population it was drawn from. There are two key components to good sample design: (a) knowing the size and distribution of relevant features of the base population (so results can be put into the context of related studies) and (b) picking the right subjects out of the base population (proportional or true random sampling).

(ii) Observational Error

Observational errors raise the question: "How can we make sure that the results from our survey actually describe the attributes in question, not some yet-unknown influence?"

Observational errors can occur as a result of expected random variation, but also through measurement bias. A typical case of measurement bias would be the wording of questionnaire items.

The compensation of observational errors poses the largest problem. A precise knowledge of the studied features, control for known confounding variables, pre-

---

[1] See for example Harrè and Secord (1972), Popper (1978), Copi (1979).

[2] For a comprehensive discussion of the critical rationalistic approach to science cf. Popper (2002b) and Popper (1972). For an evaluation of critical rationalism cf. Keuth (2002).

[3] See for example Planck (1975) and Kantorovixh (1993).

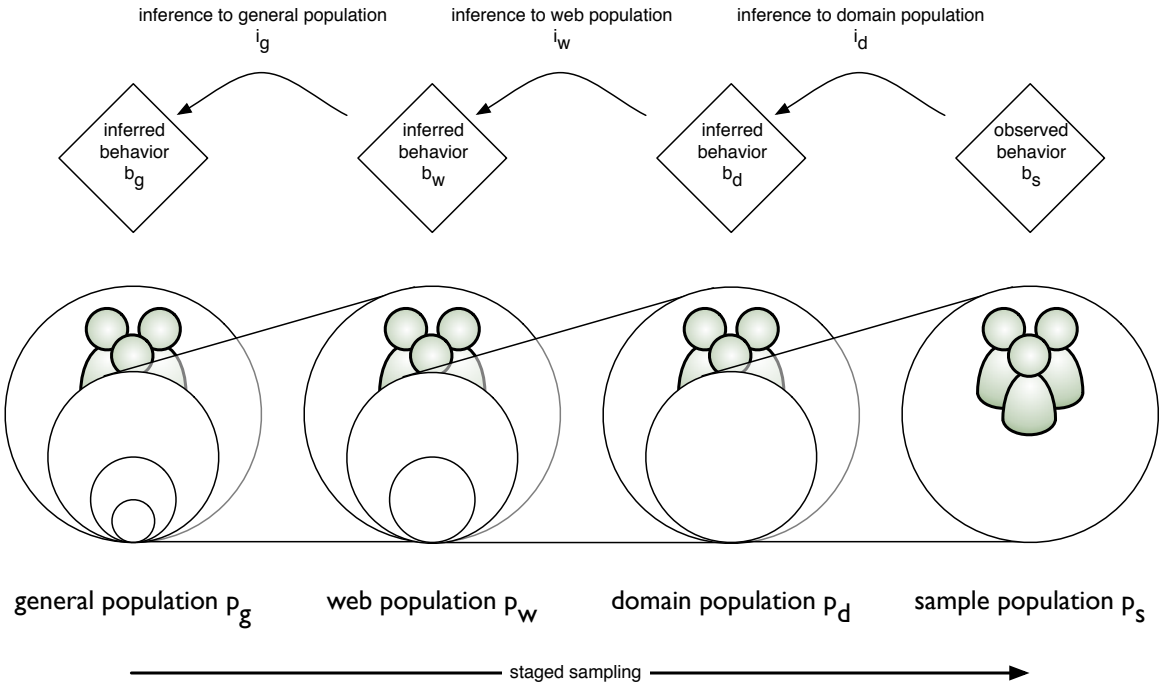

figure 1

tests of the methods are crucial components of sound research designs with minimal observational errors.

(iii) Inferential error

Inferential errors raise the question: "How can we make sure that the result from our inductive method are accurate and free from false positives (type II errors)?"

Inferential errors, which in social science were once restricted to the process of manual interpretation by the researcher, have recently invaded the methodology itself. Explorative, inductive methods of hypothesis (even insight) generation can have a bias of their own. Such risk of spurious results is most pronounced in very large data sets, as found in web science.

There are several ways of limiting inferential errors, the most effective one being to use different data sets for hypothesis generation and hypothesis testing. Others include cross-validation and permutation methods.[4]

Based on these three liabilities, we can develop a generalized methodological model, comprising several staged selection (sampling) actions (see figure 1). Each stage imposes additional conditions on the validity of inferences.

Studies in web science are expected to traverse at least six stages in all but the most exceptional cases. From the general population (p_g), only individuals utilizing the web (p_w) are considered. Within those, a certain domain of interest - such as a web site or all sites of a kind - (p_d) are selected for further investigation. Measurements are conducted upon a sample (p_s) of (p_d), producing data on the measured behavior of the sample (b_s). This data is consequently assumed to accurately describe the domain population (p_d).

Successful documentation, a representative sample and cross-validated methodology of scientific studies allow for their use in the systematic testing of hypotheses and thus the incorporation of a study in the larger scientific context of a theory. If research, on the other hand, fails to properly address sampling, observational errors and inferential errors, it withdraws itself from scientific scrutiny,encumbers peer review, and thus fails to add value to its discipline.

The Challenges and how to address them

These liabilities described above in their general form arise for web science in specific contexts. The challenges they pose, shall be discussed in the following.

Challenge A: Necessary context

As the web is increasingly permeating everyday life, more and more time is spent online, engaging in ever more meaningful transactions, dialogue and collective action. Evolving from mostly a postal mail replacement, the internet now harbors economic transactions, job markets, scientific discussion, self-help groups and so on - almost the whole spectrum of social activities can be found on the net. The advent of the long-heralded mobile web and "pervasive computing" promises to further blur the once clear cut line between human behavior on- and off-line. This trend is important to the social sciences since human behavior on-line is coded and stored by default. The advantages as compared to conventional experiments, surveys and content analyses are clear: Data on the web is potentially more comprehensive and easier to harvest than data collected off-line. Since data generation requires no change to the usual modus operandi of web usage, it can even surpass off-line data with regard to measurement errors.

Even so, the emerging field of web science has to address specific challenges that arise in spite of this data

[4] See for example White (2000), Jensen and Cohen (2000).

wealth. The data we are increasingly able to use does only document the part of human behavior that happens on and with the web. Human behavior on the web, however, is only a part of the whole of human behavior. Just as sociological theories on workplace behavior usually cannot be applied to family situations, findings from web data analysis might not apply off-line. Even within the web, there are distinct domains within which actions are governed by a specific ruleset. This has important implications.

Consider, for example, a graph of social connections from one individual. The entire graph might span a large variety of communication modalities, such as face-to-face, email, telephone, diverse social networking sites, instant messages &c. Thus, originating from its nexus are subsections that seem disconnected, even though they may overlap. Since in practice (for technical as well as ethical reasons) it is virtually impossible to obtain data on the entire graph, any observation about on-line parts of the graph must necessarily remain incomplete. Conclusions based on such analysis are severely limited in their explanatory power beyond the domain of the original data set.

To interpret data of on-line behavior correctly, scientists need an understanding of the relationship between human behavior on the whole and the part of human behavior that happens on-line. To do this, hypotheses based on established theories of human behavior have to be tested on the data collected in on-line environments. This allows the incorporation of domain specific studies of web science into the larger context of social sciences.

The challenge A: represents a sampling problem within the staged selection model of web science discussed above.

Challenge B: The dictum of culture and design

Behavior follows design and culture. Given the absence of any other external constraint (such as scarcity of resources), social forces are what ultimately shapes human behavior. In modern societies, we often speak of social situations that impose varying social codes on human behavior. This is true for human behavior off- and on-line. Guests at a dinner party converse differently from guests at a diner, although in both cases it is human interaction between strangers. Analogous codes govern the myriads of domains on the web. Some are implicit, some explicit; some are vigorously enforced by members of the community (such as etiquette rules in fora), some are not sanctioned at all.

Social codes represent important determinants of behavior, and as such have to be treated as confounding variables that must be controlled. Failure to do so can result in severe observational errors. One consequent prerequisite for studies in web science is therefore the knowledge of social codes within all investigated domains. Special care has to be taken with data sets that span domains with differing codes, since this might account for potential differences within the sample.

Social conventions can explain why a conversation through a microblogging platform is different from a conversation in the commenting section of a blog. Still, some of the attributes of messages are technologically determined (especially size, use of links &c). This is yet another influence on the data on behavior on-line. The technological design of such communication systems itself imposes binding constraints on human behavior. While statements in a microblogging conversation have to remain limited to a few lines, a comment on a blog has no such restriction. The technological boundaries on interaction serve as a filter which codes human behavior by only accepting a limited set of expressions[5]. Although web users will usually try to convey their intent as verbatim as possible, the coding process often introduces ambiguity or even misrepresents the original input. At the same time, cultural conventions can easily bend and circumvent the original design, rendering an intimate knowledge of the way people use even a known system indispensable.[6]

In order to reach meaningful conclusions about human behavior on-line, understanding and knowledge of the technological and cultural constraints that shaped the data are compulsory. Research that is aware of this challenge can support its findings through cross-validation and comparison with equivalent social scientific theories from off-line interaction.[7]

The challenge B: represents an observational bias problem within the above discussed staged selection model of web science.

Challenge C: Deceptive size of data sets

As established above, the field of Web Science is rich in data traces of human behavior. The traces form large data sets which significantly facilitate both hypothesis testing and exploratory data analysis. At a first glance, this abundance of coded information is everything a social science researcher could wish for. There are several key advantages to web data sets: (i) a potentially increased validity through larger samples (ii) a potentially increased validity through less biased measurement methods (perpetual invisible measurements during ordinary use result in high external validity) (iii) enhanced comparability of studies through common data standards, deterministic, algorithmic methods of analysis and agreed methods of comparison. .

Contrary to expectations raised by these advantages, the size of such data sets amplifies some risks. As a result of the size and technical genesis of such data sets, they are rich in statistically discernible patterns. While some patterns carry obvious explanatory power, others may only distract scientific inquiry, hide confounding factors, or be stochastic artifacts.

A “pathology” associated with the analysis of large data sets is that it often employs algorithms in

---

[5] This reduces general fidelity, and is the reason for Couper (2001) to command extra caution with data from the web.

[6] For example, think of the different interpretations of a happy emoticon in online communication: Depending on previous messages and context it can signify happiness (an internal emotional state of the sender), malicious joy, agreement &c.

[7] For a comparison of the reliability of on-line and off-line surveys, see for example Fricker (2005).

order to search for causal relations.As Jensen & Cohen 2000 show, there are three ways in which data mining algorithms can suffer from large data sets.

(i) an algorithm that assumes causal relations simply because there are many features in the data (some of which are bound to be significant by chance if the data set is large enough) commits **overfitting**. Bonferroni checks adjust the threshold of significance according to the number of features, and alleviate this problem.

(ii) an algorithm that puts special weight on certain types of data (such as preferring variables with many values (gender, race) over variables with few values (sex)) commits **attribute selection errors**. These errors are best avoided by modifying the algorithm to remove the attribute selection bias.

(iii) an algorithm whose precision degrades with growing data sets is committing **oversearching**. Such behavior can be compensated by taking the size of the data set into account. [8]

The cause of this pathology is ultimately not the algorithm, but rather the research design itself. Although explicitly discouraged, many studies use the same data set for hypothesis generation and for hypothesis validation. There are several motives for this; in some fields such as economics, it is hard to do otherwise, in others, such as web science, the availability of data favors such measures. Nevertheless, such research engages in the hazardous practice of "data snooping". Without theoretical guidance and testable hypotheses, significant correlations can always be found for large enough data sets, and spurious patterns (such as patterns stemming from certain cyclical data acquisition practices) are easily misinterpreted as meaningful findings.

A typical example where these pathologies have a severe impact on research is economical time-series analysis. In that field, "there is little explicit guidance from theory regarding the identity of the predictive variables".[9] As a result, over-searching and over-fitting of models produce predictive models that appear to be valid. The successful buying and selling of stock seems to follow a clear date rhythm. In at closer examination of calendar effects, however, Sullivan, Timmermann and White find that none of the models remains significant when data snooping effects are compensated[10].

The challenge C: represents an inferential problem within the above discussed staged selection model of web science.

Challenge D: Accountability in unknown universes

It has already been made clear that sound sampling is a core issue for successful web science. Nevertheless, there is an aspect to it that has special significance for this area of research. Many of the proposed best practices above rely on the researcher's ability to assess the quality of the sample, and more specifically its representativity.

Representativity, however, is defined as the correspondence of feature distributions within a population and a respective sample. In order to detect a possible bias in assumed feature distributions, at least a realistic estimate of the total population is required. In the case of web science, this poses a great challenge. Not only are essential feature distributions (such as age, education, nationality &c - all of which are basic features well-covered by off-line sociology) unknown for many domains. Even the size of populations can be difficult to assess.

Access statistics of web sites, for example, may log site visits and utilize tracking methods in order to identify individuals. Even so, the true number of unique visitors cannot ultimately be measured: If multiple people use the same computer, they are collapsed into one identity . If one person utilizes several different computers, she is split into separate identities.

An even greater challenge can lie in the exact definition of domains within the web. A population of interest must be described by an unambiguous rule that clearly accepts or rejects individuals. For example, a definition such as "all users that visited site s in march 2008" satisfies the condition. As a contrast, a definition such as "all users of social networking sites in march 2008" is not acceptable, since "social networking sites" is an ambiguous term that requires further description.

The challenge of unknown statistical universes can be addressed through several methods. Most importantly, basic socio-demographic data about domains on the web has to be collected and published. Secondly, since a complete census is unfeasible, either good random samples or stratified samples have to be used instead of mere opportunity samples. This ensures a coherence between the sample and the original population. Additionally, multiple samples may be drawn and cross-validated. These procedures help to ensure valid sampling, and thus support the validity of web science studies.

The challenge D: represents a sampling problem within the above discussed staged selection model of web science.

A Blueprint for Research

Even if all of the challenges described above are competently addressed by a study, it still has to transcend its isolated state and establish a connection to the established scientific literature. To avoid research that is simply an aggregation of isolated and maybe even misleading observations, the authors propose a methodological approach to Web Science that incorporates theory in research practice. This approach is

---

[8] For detailed discussion of the respective problems, please see White (2000), Sullivan, Timmermann and White (1998), Cooper & Gulen (2006), Jensen & Cohen (2000).

[9] cf. Cooper & Gulen (2006) 1264.

[10] cf. Sullivan, Timmermann and White (1998).

based on the critical rationalistic view of science as formulated in the scientific method.[11]

(1) Identify a specific area of interest in the field of Web Science;
(2) Formulate a theory on the processes that rule this field of interest, based on intuition, preliminary data analysis or analogue processes in related scientific fields;
(3) Derive from this theory hypotheses which can be tested on available data sets;
(4) Identify and prepare an appropriate data set to test the hypotheses of (3);
(5) Test the hypotheses of (3) on the data set of (4). If the hypotheses do not hold true start again with (2) and incorporate the new understanding of the topic. If the hypotheses are corroborated go to (4) and identify a new data set which could be used to test the hypotheses further.

Conclusion

As described above, web science is no trivial field of scientific discovery. The prospective researcher is faced by technological and theoretical challenges that need to be addressed. This endeavor is further complicated by the scarcity of best practices for web science research, as well as the absence of a sound, agreed upon methodology for the field. A methodological discussion about the most promising approach to Web Science could prove to be very fruitful at this early stage in the development of the field. The authors understand this paper as an invitation to start such a discussion. Until addressed, publications with severe sampling issues and biases such as (Wadhwa et al. 2009) remain a burden on this young field.

Acknowledgements

The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

Copi, I. M. (1979). Symbolic logic. Prentice Hall.

Cooper, M., Gulen, H. (2006). Is Time-Series-Based Predictability Evident in Real Time? Journal of Business, 79(31):1263-1292.

Couper, M. P., Traugott, M. W., and Lamias, M. J. (2001). Web survey design and administration. The Public Opinion Quarterly, 65(2):230–253.

Fricker, S., Galesic, M., Tourangeau, R., and Yan, T. (2005). An experimental comparison of web and telephone surveys. The Public Opinion Quarterly, pages 370–392.

Godfrey-Smith, P. (2003). Theory and Reality: An Introduction to the philosophy of science. The University of Chicago Press.

Harrè, R. and Secord, P. F. (1972). The explanation of social behavior. Blackkwell.

Jain, A. K., Dhuin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1):4–37.

Jensen, D. D., Cohen, P. R. (2000). Multiple Comparisons in Induction Algorithms. Machine Learning, 38(3):309-338.

Kantorovixh, A. (1993). Scientific discovery: Logic and tinkering. State University of New York Press.

Keuth, H. (2002). Was bleibt vom kritischen Rationalismus. In Böhm, J. M., Holweg, H., and Hoock, C., editors, Karl Poppers kritischer Rationalismus heute: zur Aktualität kritisch rationaler Wissenschaftstheorie, 43–57. Mohr Siebeck.

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., and Couper, M. (2004). Pyschological research online: Report of board of scientific affairs adivosry group on the conduct of research on the internet. American Psychologist, 59:105–117.

Nosek, B. A., Banaji, M., and Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. Group dynamics, 6(1):101–115.

Orlikowski, W. J. (2000). Using technology and constituting structures: a practice lens for studying technology of organizations. Organizational science, 11(4):404–428.

Planck, M. (1975). Positivismus und reale Außenwelt. In Vorträge und Erinnerungen, 228–249. Wissenschaftliche Buchgesellschaft Darmstadt.

Popper, K. (1972). Objective Knowledge: An Evolution Approach. Oxford University Press.

Popper, K. (1978). The unity of method. In Bynner, J. and Stribley, K. M., editors, Social research: Principles and procedures, 17–24. The Open University.

Popper, K. (2002a). Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge.

Popper, K. (2002b). The Logic of Scientific Discovery. Routledge.

Skitka, L. J. and Sargis, E. G. (2005). Social psychological research and the internet: the promise and the perils of a new methodological frontier. In Amichai-Hamburger, Y., editor, The social net: the social psychology of the internet. Oxford University Press.

Sullivan, R., Timmermann, A., White, H. (1998). Dangers of Data-Driven Inference: The Case of Calendar Effects in Stock Returns. UC San Diego Department of Economics Discussion Paper 98-31.

Thelwall, M. (2002). Methodologies for crawler based web surveys. Internet Research: Electronic Networking Applications and Policy, 12(2):124–138.

Wadhwa, V., Saxenian, A., Freeman, R., Gereffi, G., Salkever, A., (2009). America's Loss is the World's Gain: America's New Immigrant Entrepreneurs, Part IV. Working Paper Available at SSRN: http://ssrn.com/abstract=1348616

White, H. (2000). A Reality Check for Data Snooping. Econometria, 68(5):1097-1126.

[11] For a discuission of the relationship between critical rationalism and the scientific method cf. Godfrey-Smith (2003) 69-72.