Challenges and Opportunities for Computational Social Science

Strohmaier, Markus; Zens, Maria

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

6.1 Challenges and Opportunities for Computational Social Science

Markus Strohmaier, Maria Zens (GESIS)

The field we would like to sketch out – Computational Social Science (CSS) – is an emerging area of research situated at the intersection of Social and Computer Sciences. According to the Computational Social Science Society of the Americas (CSSSA), the subject matter can be outlined as "The science that investigates social phenomena through the medium of computing and related advanced information processing technologies." The two-fold orientation of CSS towards algorithms and Social Sciences might prove beneficial for both disciplines. On the one hand, CSS reaches out to offer means for processing large amounts of data to the Social Sciences and, on the other hand, takes hypotheses and theories from the Social Sciences to arrive at meaningful models of social behavior which can be applied to and tested against large data sets taken, for example from social media.

Data-induced opportunities and challenges for CSS

The formation of CSS responds to a situation in which interactions in the digital world generate and shape social structures in a novel way and, in doing so, provide social research with prolific new data sources. The increasing integration of the World Wide Web in our daily lives already has created massive volumes of social data, i.e. data about humans' everyday behavior and social interactions in the real world. Such social data opens up exciting new opportunities as well as challenges for computer and social scientists to work towards a new and deeper quantitative understanding of complex social systems. At the same time, the increasing availability of such social data has led to new types of and directions for research. In our contribution to this volume, we will discuss these and other related issues from both a Computer Science and a Social Sciences perspective. We will give examples from recent and current research that illustrate how the use of new and large data sets can support the Social Sciences in analyzing socio-political phenomena such as mobility issues, interpersonal communication, and the structures of political discourse.

Currently, the Social Sciences appear to be a discipline in distress; as the sociologists Mike Savage and Roger Burrows put it some years ago: "both the sample survey and the in-depth interview are increasingly dated research methods, which are unlikely to provide a robust base for the jurisdiction of empirical sociologists in coming decades" (Savage and Burrows 2007: 885). The possible turn-out of this "crisis" is that we are about to encounter a data-driven advancement of a well-established discipline and an ever deepening and serious collaborative intent on all sides.

The confluence of Social Sciences and Computer Science seems natural in a situation in which both sides are in need: computer scientists need to make social sense of "big data" and social scientists require tools to handle new amounts (and the new quality) of data that go beyond their traditional ways of collecting, structuring, and evaluating.

Having stressed the challenges in quantity and the social novelties posed by transactions and communicative interactions on the web, one has to bear in mind that data-driven progress is nothing new in the history of the Social Sciences. A brief detour into history might illustrate that, although the kinds of data might be new, the fact of fruitful co-operation is surely not.

The paradigm: Technology-driven advancement in processing social data

Herman Hollerith (1860–1929), the son of German immigrants to the US, is an early example of how technology-driven innovation at the intersection of Social Sciences and what became Computer Science can be established. In his work on "An Electric Tabulating System" (1889), which formed the basis for his PhD at Columbia one year later, Hollerith developed a crucial foundation for the advancement of Computer Science during the 20th century. Hollerith, who made use of previous technological knowledge gathered by the textile industry (Jacquard loom techniques), developed a mechanical tabulator based on punched cards to rapidly tabulate statistics from millions of pieces of data. His machines were able to tally not only overall numbers, but also individual characteristics and even cross-tabulations; he invented the first automatic card-feed mechanism and the first key punch. The prime use case for his invention was social data: Hollerith built calculating machines under contract for the US Census Office; in 1890 Hollerith machines were first used to tabulate US census data, and, subsequently, in many more censuses in various countries. Due to

his invention, processing time could be decreased enormously; while it had taken about eight years to tabulate the 1880 US census, the 1890 census using his machines took – figures differ on this point – only one year or even less. Hollerith's firm "Tabulating Machine Company" merged with others to become the "Computing Tabulating Recording Company", which was later renamed "International Business Machine Corporation" (IBM).

One might reasonably dispute the grounds for progress – whether the advancement in technology fostered efficiency for the administration or, vice versa, whether administrative needs induced the development of technology. However, with his invention Hollerith provided the means to leverage the processing of administrative data on the US society and in many more states.

Bearing this landmark in mind and being in the midst of a new deluge of digital data that we are struggling to make social sense of, we should ask what are the respective 21st century challenges and opportunities? How should "modern Hollerith machines" be structured and where could they be deployed? We will give a few examples to try and find preliminary answers to these questions.

New kinds of data on macro, micro, and meso scale: Mobility as a use case

First, we will look at the opportunities that mobility data offer the Social Sciences. Human mobility in societies is one of the key issues that can be addressed with large data-sets generated from GPS or cell-phones. We will take the three-level-approach familiar to Humanities and Social Sciences and look at the impact of new kinds of data in this field at the macro, meso, and micro levels. On the macro scale, GPS-data or check-in-data from social media applications like Foursquare, Gowalla or Twitter provide information about the locations, destinations and travel modes of people and give us a kind of "big picture" of mobility. These data are increasingly available and being used for research. Cheng et al., for instance, analyzed the use of location sharing services by investigating 22 million check-ins by 220000 users (Cheng et al. 2011). What is interesting from a Social Sciences point of view is that the authors not only studied spatio-temporal mobility patterns, but also analyzed the correlation of social status and mobility behavior. In fact, they themselves regard this aspect of their research as "one of the more exciting possibilities raised by the social structure inherent in location sharing services" (ibid.: 87).

On the meso level, the Amsterdam real time project can be mentioned. The project dates from 2002: Amsterdam residents were asked to use a tracer unit with GPS, the real-time visualization of the collected data showed lines against a black background and hereby constructed a (partial) map of Amsterdam based on the actual movements of real people. The project was carried out in conjunction with artists and became part of the exhibition "Maps of Amsterdam 1866–2000" in the Municipal Archive of Amsterdam. Almost a decade later and with advanced technologies, Calabrese et al. conducted a case study on Rome ("Real Time Rome"), which also made use of real-time mobility data and was developed for the Tenth International Architecture Exhibition of the Venice Biennale. (Calabrese et al. 2011) For this project a monitoring system using a variety of tools was deployed to grasp urban mobility from cell-phones and locational data from the public transport system. Among other issues, the density of people using mobile phones at historic sites or during events was measured and visualized. The distributions revealed dynamics of movement during the course of the day, as well as hot spots for tourists or event gatherings. The practical impacts of such services cover the optimization of urban planning, of services and traffic information, the reduction of inefficiencies and the support of efforts to put public transportation where the people need it.

For the significance of new data on the micro scale we would like to turn to an experimental project in which Radio Frequency Identification (RFID) was used to monitor the dynamics of communicative interaction between people (Cattuto et al. 2010). At the core of this group's work is the aim to present micro level interpersonal relations at high resolution, but with the clear objective to provide tools that can be scaled up. Sensing tiers with unobtrusive RFID devices were embedded into conference badges to sense face-to-face interactions and spatial proximity of participants as well as the duration of contacts. One of the results was a power law distribution with identifiable "super-connectors", the "crucial actors in defining the pattern of spreading phenomena", who "not only develop a large number of distinct interactions, but also dedicate an increasingly larger amount of time to such interactions" (ibid.: 5).

"Digital society": data sources for analyzing political discourse and social action

Shifting from the granularity of data to the topical areas relevant for the Social Sciences, one has to recall the double function of the new social media as sources of data and social arenas in their own right. In their introduction to the special issue of "Sociology" on the relationship of new technologies and society, Linda McKie and Louise Ryan point out: "New technology is not simply capturing but actively constituting social interaction"(McKie and Ryan 2012: 6). Therefore, both the structures of "digital society" and the reflections of social behavior or political discourse in the digital world are of interest.

In what can be regarded a seminal paper for Computational Social Science, Lazer et al. (2009) show a visualization of political conversations in the blogosphere around the 2004 US election and made this a prime example for how existing socio-political theories can profit from examining vast data. They displayed the network structure they found in a community of political blogs, where this structure – with a clear distinction between the liberal and the conservative camps – reflected the political map very closely.

The impact of web technologies on political events has been discussed in, for example, the context of the so-called Arab Spring, when communication through social media channels played an important role for both the diffusion of alternative political information and the organization of the protest movement. Starbird and Palen looked at the uprising in Egypt and analyzed the most influential Twitter users during the revolution in early 2011, with influence being measured by the number of retweets and followers. They found individuals, bloggers, journalists, and mainstream news channels among the most influential actors. Yet first up on the crucial days in January 2011 is the internet activist Wael Ghonim (Starbird and Palen 2012). In our own research (conducted with Lichan Hong at Xerox Parc) on political conversations on Twitter (about 100 million tweets) during the Egyptian revolution 2011, we found that the hashtag "#jan25" – which denotes the first day of the protests – was a top-trending hashtag prior to the actual date.

These findings show, that social media might serve as an additional channel for the dissemination of mainstream information, but also as alternative channels for independent journalists and activists; basically, they are an open communication space for governments, traditional media, social movements, and dissidents alike. Social media are important in the competition for political

hegemony and interpretation, which becomes evident when they are subject to regulation, censorship, and surveillance.

Finally, we would like to mention research on the representation of the German parliamentary elections 2013 in social media, which is work in progress at GESIS. We analyze twitter accounts of electoral candidates from the various parties; we try to apply models of communication taken from the Social Sciences to the conversations on Twitter in the run-up to the election and look at topical conversation practices (via hashtagging) and structural conversation practices such as mentioning and re-tweeting. In so doing, we want to investigate the similarity between parties as measured by hashtags, the foci (both topical and structural) of partisan communication and the stability of conversation practices (i.e. measure focus shifts at certain points in the electoral time-line).

Challenge: providing computational infrastructures

With these examples, we have tried to highlight both the data side of CSS – with respect to modelling on the macro, the meso, and the micro levels – and the Social Sciences side of CSS – with respect to their contribution to revealing behavioral patterns in socially and politically relevant realms. Since the digital world is tracking the social world more and more closely and specific forms of a "digital society" emerge, CSS sets out to use computation to discover patterns, build models, validate social theories and learn about societies. Further to that, we have to address data management issues, archival issues, and legal issues concerning privacy and data protection.

The efforts, of course, go beyond the ones we could mention and take various angles – data mining and processing, sociology, political science, network analysis etc. –, but the main challenge for research infrastructures is to provide computational infrastructures for dealing with (1) more data: for analyzing large amounts of data, (2) fuzzy data: for cleaning up imprecise and noisy data, (3) new kinds of data: for processing real-time sensor-streams and web data, (4) correlations: for understanding what (in addition to why).

This brings us back to Herman Hollerith and his technological achievement. To specify and build what we may call "modern Hollerith machines" is a major challenge for Computational Social Science. There already exists a range of tools and platforms for extracting big data streams (Hadoop, Mahout, SAMOA, S4, Storm, R, WEKA, MOA – cf. De Francisci Morales 2013); what is needed are

algorithms and clustering techniques that focus on social structures and expand the horizon of data modelling from space and time complexity to include social complexity.

Challenge: computation-focused social theories

From the point of view of the Social Sciences and their "crisis", the challenges are mainly attached to the large amounts of and the uncontrolled quality of the new data at stake. These data have the advantage of being easily accessible, but they do not meet the traditional standards of social science research. They are not intentionally collected under ceteris-paribus-conditions, but "found data". They are often far from being representative, and suffer from single channel and self-selection biases. In short: using transaction data from social media for Social Sciences research requires new, robust methods of data collection, cleansing and evaluation. Moreover, Social Scientists should take on the challenge to further include computation-focused social theories, e.g. network theories.

Opportunity: CSS – more than adding up expertise

The confluence of Social Sciences and Computer Science merges expertise: Computer Science offers the ability to process large data sets and provides algorithms and methods of data mining. The Social Sciences contribute their knowledge of social theories, methods, data collection, and relevant issues. This means more than simply adding up the things that we knew before. While working together on Computational Social Science issues, both knowledge systems are being transformed by the opportunities of analyzing new amounts of digital information with regard to social systems.

References

Calabrese, F./Colonna, M./Lovisolo, P./Parata, D. and Ratti, C. (2011): Real-Time Urban Monitoring Using Cellular Phones: A Case-Study in Rome. IEEE Transactions on Intelligent Transportation Systems 12 (1), 141–151.

Cattuto, C./Van den Broeck, W./Barrat, A./Colizza, V./Pinton, J. F. and Vespignani, A. (2010): Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. PloS One 5 (7), e11596.

Cheng, Z./Caverlee, J./Lee, K. and Sui, D. Z. (2011): Exploring Millions of Footprints in Location Sharing Services. ICWSM, 81–88.

De Francisci Morales, G. (2013): SAMOA – A platform for mining big data streams. Proceedings of the 22nd international conference on World Wide Web companion, 777–778.

Hollerith, H. (1889): An Electric Tabulating System. The Quarterly, Columbia University School of Mines, Vol. X , No. 16, 238–255.

Hollerith, H. (1890): In connection with the electric tabulation system which has been adopted by U.S. government for the work of the census bureau. Ph.D. dissertation, Columbia University School of Mines.

Hollerith, H. (1894): The Electric Tabulating Machine. Journal of the Royal Statistical Association 57 (4), 678–682.

Lazer, D./Pentland, A. S./Adamic, L./Aral, S./Barabasi, A. L./Brewer, D., [...] and Van Alstyne, M. (2009): Life in the network: The coming age of computational social science. Science 323 (5915), 721–723.

McKie, L. and Ryan, L. (2012): Exploring Trends and Challenges in Sociological Research, Sociology 46 (6), 1–7.

Savage, M. and Burrows, R. (2007): The coming crisis of empirical sociology. Sociology 41 (5), 885–899.

Starbird, K. and Palen, L. (2012): (How) will the revolution be retweeted? Information diffusion and the 2011 Egyptian uprising. Proceedings of the ACM 2012 conference on computer supported cooperative work, 7–16.