Commentary

BIG DATA & SOCIETY

Big Data & Society
July–December 2015: 1–6

DOI: 10.1177/2053951715617185
bds.sagepub.com

SAGE

Small decisions with big impact on data analytics

Jana Diesner

Abstract

Big social data have enabled new opportunities for evaluating the applicability of social science theories that were formulated decades ago and were often based on small- to medium-sized samples. Big Data coupled with powerful computing has the potential to replace the statistical practice of sampling and estimating effects by measuring phenomena based on full populations. Preparing these data for analysis and conducting analytics involves a plethora of decisions, some of which are already embedded in previously collected data and built tools. These decisions refer to the recording, indexing and representation of data and the settings for analysis methods. While these choices can have tremendous impact on research outcomes, they are not often obvious, not considered or not being made explicit. Consequently, our awareness and understanding of the impact of these decisions on analysis results and derived implications are highly underdeveloped. This might be attributable to occasional high levels of over-confidence in computational solutions as well as the possible yet questionable assumption that Big Data can wash out minor data quality issues, among other reasons. This article provides examples for how to address this issue. It argues that checking, ensuring and validating the quality of big social data and related auxiliary material is a key ingredient for empowering users to gain reliable insights from their work. Scrutinizing data for accuracy issues, systematically fixing them and diligently documenting these processes can have another positive side effect: Closely interacting with the data, thereby forcing ourselves to understand their idiosyncrasies and patterns, can help us to move from being able to precisely model and formally describe effects in society to also understand and explain them.

Keywords

Validation, evaluation, error analysis, data preprocessing, social network analysis, data mining

Introduction

Big social data have enabled new opportunities for evaluating the applicability of social science theories that were formulated decades ago and often based on small- to medium-sized samples in today's contexts and for social agents operating in contemporary socio-technical infrastructures. These data, which includes large-scale traces of social interactions and natural language use, are also essential for developing new knowledge and methods based on bigger and broader datasets than those typically used in the past (Lazer et al., 2009).

In some cases, Big Data coupled with powerful computing have the potential to replace the statistical practice of sampling and estimating effects by measuring phenomena based on full populations. For instance, the social networks concept of small worlds, which basically means that a randomly picked pair of people is linked through a small number of intermediaries or social circles, is based on a few experiments where, for example, 64 (29%) out of 296 chain letters successfully arrived at their predefined target (i.e. a person unknown to the first sender); with a median of 5.2 (N about 16) intermediaries (Travers and Milgram, 1969). This study has recently been replicated based on Facebook data (a graph with about 721 million nodes), using a bird's eye view (automated efficient graph search) rather than a frog's perspective (local search) (Backstrom et al., 2012). An average of 3.7 intermediaries was found for the Facebook data.

UIUC, USA

Corresponding author:
Jana Diesner, UIUC, 501 E Daniel Str, Champaign, IL 61820, USA.
Email: jdiesner@illinois.edu

This shorter distance might be due to the fact that the participants in Milgram's studies had incomplete knowledge about chains of acquaintances beyond their ego-network or friend-of-a-friend network (1.5° to 2° of separation), while in the Facebook study, algorithms performed the search task. An alternative explanation would be that the average social distance has decreased over time, e.g., due to the wide diffusion of communication and social interaction technologies. Has our social world truly become smaller over the last 50 years? Finding and proofing reasons and explanations for such empirical observations require more work. Another example comes from political science, where scholars have been collaborating over decades to define and update categorization schemas for geopolitical actors and events with the goal of enabling the analysis of international relations, especially conflicts (Schrodt et al., 2004, 2008). Starting from the same underlying approach, the "Global Database of Events, Language and Tone" (GDELT)[1] serves the same purpose but is based on big event data: Enabled by computing infrastructure providers, a large number of national and international news content providers, and research in information science, GDELT provides a continuously updated geopolitical event database with over a quarter-billion events from 1979 onward (Gao et al., 2013).

In the (computational) social sciences and (digital) humanities, researchers have long started to use readily available technologies, including APIs, computing environments like R, and scripting languages like Python, to enhance methods common in their fields, e.g. close reading and text coding, as well as advanced social data analytics and visualization techniques, such as topic modeling, sentiment analysis, and clustering (Abello et al., 2012; Talley et al., 2011; Underwood et al., 2013; Wang et al., 2007). These improved methods are then applied to big social and cultural data.

The outlined leap forward in supporting a better understanding of society, culture, and socio-technical systems at scale benefits from a combination of developments: Electronic repositories and collections of data (e.g. the Stanford Large Network Dataset Collection[2]), code (e.g. GitHub[3]), scientific publications (e.g. DBLP[4]), and user-generated (e.g. Wikipedia[5]) as well as traditional print content (e.g. HATHI Trust[6]) support the reproducibility of findings based on Big Data and the sharing of material—at least theoretically (Cragin et al., 2010; Lane et al., 2014). Innovations to copyright law, such as the Creative Commons Licenses,[7] which apply to Wikipedia data for instance, and open source software licenses, such as the Apache License[8] or the GNU General Public Licenses,[9] further ease the process of sharing and reusing information and tools. Also, some commercial providers, platforms, and websites that receive, manage, and synthesize social media content offer APIs that allow researchers and practitioners to access and analyze big social data, e.g. Facebook,[10] Twitter,[11] and Yelp.[12]

So what's the problem?

Preparing big social data for analysis and conducting actual analytics involves a plethora of decisions, some of which are already embedded in previously collected data and built tools (Diesner, 2013; Moore et al., 2000). These decisions refer to the recording, indexing, and representation of data and to the settings and (parametric) choices for analysis methods. For example, when fusing data from various social media sites ("v" for variety of Big Data), one needs to think about how to identify identical users across different platforms—where the same person might use different names on different sites or multiple names on the same site, and different people might use the same name on different sites—and what errors in resolving these ambiguities mean for the accuracy of the data and obtained findings (Iofciu et al., 2011; Zafarani and Liu, 2013). Means for communicating and learning about these decisions include data annotation, such as meta-data, and diligent documentation. In the given disambiguation example, one might need to dig deep into meta-data, individual language use, or interaction patterns, to tell users apart or merging them. Another instance of possible issues with reusing Big Data is the considerable amounts of false positives that have been found in GDELT, where these problems result from the same event being reported by multiple sources and difficulties with automatically disambiguating these redundancies (Hammond and Weidmann, 2014).

While such decisions can have tremendous impact on research outcomes, they are often not obvious, not considered, or not made explicit (De Choudhury et al., 2010; Howison et al., 2011; Ruths and Pfeffer, 2014). Consequently, our awareness and understanding of the impact of these decisions on analysis results and derived implications are highly underdeveloped. This might be attributable to occasional high levels of over-confidence in computational solutions as well as the possible yet questionable assumption that Big Data might wash out minor data quality issues, among other reasons. Ultimately, it is up to the users of big social data to leverage given resources in a responsible and meaningful way and to bring relevant questions and appropriate analysis techniques to the data. More research on the quality of big social data can aid this process.

In our work, we have begun to address this issue by identifying the quantitative and qualitative impact of small decisions made prior to and throughout the research process on its outcomes. More specifically, we have been measuring the effect of inaccuracies in

resolving entities when constructing social network data and of refining previously built lexical resources on analysis results and the interpretation of findings (Diesner, 2013, 2015; Diesner and Evans, 2015; Diesner et al., 2015; Kim and Diesner, 2015a, 2015b). The next section provides one example for each type of research.

Before turning to this main point, I am concluding the problem statement with the brief mention of a related issue that is only tangential to this article and will therefore not be further addressed herein: One challenge that scholars seeking to use big social data need to face is that the terms of service and intellectual property/copyright regulations for publicly accessible data, especially those synthesized, organized, and hosted by commercial providers, can put a damper on educational and research aspirations. The fact that large-scale interaction and language use data are visible and available to the public does not necessarily mean that researchers are also allowed to collect and analyze them. APIs can help to mitigate this issue. The relationship between the legality, feasibility, and ethics of data acquisition and analytics is an evolving aspect of big social data science. Discussing related policies and regulations might be another essential ingredient for making progress with practical and computational solutions.

Scrutinizing the impact of data quality issues on social computing research

Ambiguity of social network data

Entity resolution involves two tasks: First, locating and consolidating the different references to a single unique entity. This applies, for example, to (a) social media networks, where individuals might use different names on different platforms; (b) social networks constructed from text data, where people and organizations might be referred to by various ways of expressing their names and roles (John Kerry, United States Secretary of State) and pronouns; and (c) scientific collaboration and cocitation networks, where authors might be indexed with different variations of their names, e.g., with and without middle name initials. Multiple computational, algorithmic, and human-in-the-loop solutions have been developed to address these problems. For example, the ORCID project[13] was started to resolve ambiguous names of authors of scientific publications relying on the input and data verification from scholars. Other providers of information about scientific publications and citations use highly accurate algorithmic solutions, sometimes coupled with the manual resolution of ambiguous cases, e.g. DBLP. The second entity resolution task is splitting up nodes that represent multiple distinct entities that are referred to by the same name. This can happen, for instance, when some people's names entail common first and last names. For example, the University of Michigan has two established scholars with the name of Mark Newman: a physicist who studies networks[14] and an HCI scholar[15] (for the curious reader: the http://howmanyofme.com/ webpage gives an idea of how many US-based people share a certain first name/last name combination). Telling both Dr Newmans apart requires knowledge about their middle names, details about their institutional affiliation (i.e. meta-data), or contextual information about their work. Prior work has shown that the problem of telling people with identical names apart is particularly important when working with Asian names (Zhao and Strotmann, 2011). Overall, prior research on entity resolution has resulted in highly accurate and automated techniques to both the consolidation and splitting of names (see, for example, Fegley and Torvik, 2013).

We have been bringing the following question to this problem: How much does entity resolution matter for big (and small) social network analysis? Our results suggest that commonly reported network metrics, as well as derived implications, can strongly deviate from the truth—as established based on ground truth/gold standard data or approximations thereof—depending on the efforts dedicated to the data preprocessing step of entity resolution (Diesner et al., 2015). We found the identification of key players to be less sensitive to entity resolution errors than variations in network metrics. For working with email data, our results have shown that failing to consolidate email addresses (i.e. indexing all email addresses that a person uses as one node) can make email networks appear less coherent and integrated and also bigger than they really are, potentially suggesting a false need for more coordination and communication. For copublishing networks, failing to split up nodes that represent multiple individuals with the same name can make scientific communities look denser and more cohesive than they are, and make individual authors appear more productive, collaborative, and diversified than truth has it, potentially downgrading the need for (interdisciplinary) collaboration and funding. We observed that in coauthorship networks, incorrect or skipped entity resolution can even lead to the misidentification of applicable network topologies, e.g. detecting power law distributions of node degrees and assuming an underlying preferential attachment process where there is insufficient empirical evidence for this claim (Kim and Diesner, 2015a, 2015b).

Scope of auxiliary lexical resources

The Big Data wave also eased access to sizable auxiliary material, which can help to disambiguate and

contextualize information, among other uses.[16] For instance, Wikipedia offers additional information on a large variety of social entities, e.g. sociodemographic meta-data on individuals and product portfolios for companies. Also, Freebase,[17] which inherited its content from Metaweb and later fed into Google's knowledge graph, used to be a big repository of structured and categorized information about a large variety of types of entities and phenomena. Freebase obtained its content from various sources, including user-generated contributions.[18] Another example is WordNet,[19] which groups English terms into about 117,000 sets of synonymous words and provides a database of relationships between words, including hyperonyms (super-subordinate relations), meronyms (part-whole relations), and antonymous adjectives (Fellbaum, 1998, 2005). The service has since been provided for many other languages as well, including Afrikaans, Persian, and Latin.[20] Similarly, a plethora of research initiatives has provided dictionaries, i.e. registers of word-category pairs (occasionally enriched with additional information such as parts of speech), that can be used to assess, for example, the subjectivity, emotionality, honesty, and morale of (pieces of) text data or their authors (Graham et al., 2009; Tausczik and Pennebaker, 2010; Wilson et al., 2005). Using such resources is not only efficient, but also a scientifically solid strategy as many of these helper tools have been previously validated—by experts or crowds of ordinary people—and documented. Moreover, this general approach puts the idea of sharing and reuse into action. However, when leveraging readily available material for Big Data analytics, we have a lack of understanding, ground truth data, and pertinent benchmark results for how much adjustment of preexisting resources to a given dataset, domain, or time period is needed in order to obtain reliable and comprehensive results.

We have started to address this issue by asking this question: How much of a difference does the tuning of lexical auxiliary material make for text mining projects? To give an example, we have used a previously built and widely adopted subjectivity lexicon (Wilson et al., 2005) in order to identify the emotionality and sentiment of information exchanged via emails. Our overall purpose with this work was to develop a novel method for efficiently assessing structural balance in large-scale communication networks (Diesner and Evans, 2015). This idea was motivated by the aforementioned opportunity and need to test the validity of social science theories in today's contexts and interaction environments; a precondition for advancing theory and substantive knowledge about social networks. This bigger goal involves small decisions throughout the research process: The original subjectivity lexicon that we used was built based on world press data (Wiebe et al., 2005), while the data for our study were email conversations from a company, namely Enron. The original lexicon associates syntactically disambiguated (via parts of speech) terms as well as their stemmed versions with a value for polarity (positive, negative, or neutral) and strength of polarity (weakly or strongly subjective). Not knowing to what degree this lexicon would generalize to the business domain, we identified and corrected for false negatives (prevalent subjective terms contained in the email data but not in the lexicon) and false positives (subjective terms included in the lexicon that were a misfit for our data). In our example, correcting for false negatives involved the detection of salient terms, such as words with a high weighted term frequency, inspecting them one by one to decide whether they should be added to the lexicon, and if so deciding on the best fitting part of speech, polarity, and strength. Removing false positives from the lexicon meant to compute a list of lexicon terms that frequently occurred in the email data, inspecting them, and modifying their values or dropping them from the lexicon. Overall, we added 34 terms (a tiny fraction, the original lexicon contains over 8200 entries), dropped 591 terms, and modified 30 entries. Overall, we changed less than 8% of the original dictionary. Even though computer-assisted, this process is tedious. Is it worth it? In our example, adjusting a lexical resource to a different domain and corpus leads to similar overall findings about balance as with using the original lexicon, but resulted in more empirical evidence (we annotated 11.9% or about 17.5K more triads with sentiment values) and lower statistical variance of the results (1.84% instead of 2.88%). We argue that checking and correcting for false positives (false alarms) and false negatives (blind spots) should be common practice when reusing lexical resources for big social data analytics. More research is needed to identify best practices, stopping criteria (how much is enough?), and evaluation procedures and metrics for this step.

Conclusions

In summary, this article argues that checking, ensuring, and validating the quality of big social data and related auxiliary material is a key ingredient for empowering users to gain reliable insights from these data. This main point is further substantiated by the fact that oftentimes, assessing the accuracy and validity of Big Data and respective findings is difficult to infeasible. This is due to population size ("v" for volume in Big Data) and the continuously evolving and changing nature ("v" for velocity of Big Data) of social systems. Furthermore, when working with large-scale digital trace data, such as social interaction and information dissemination data from Facebook and Twitter, it is

common practice not to interact with the user population, but rather to harvest and mine the information they produce and leave behind, and occasionally infer or predict additional information based on that. Ethical standards around this process keep being discussed (Kosinski et al., 2015). All of these characteristics make the collection of gold-standard or ground-truth data and the creation of benchmarks for validation a daunting and costly task.

Scrutinizing big social data for accuracy and integrity issues, systematically fixing them, and diligently documenting these processes can have another positive side effect beyond boosting the reliability of results and the reusability of material: Closely interacting with the data, thereby forcing ourselves to understand their idiosyncrasies and patterns and learning about the content domain, can help to move us from being able to precisely model and formally describe effects in society to also understand and explain them.

Acknowledgement

I thank Craig Evans and Jinseok Kim, both from UIUC, for their substantial contributions to the referred research.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported in part by KISTI (Korea Institute of Science and Technology Information).

Notes

1. http://gdeltproject.org/
2. https://snap.stanford.edu/data/
3. https://github.com/
4. http://dblp.uni-trier.de/
5. https://en.wikipedia.org
6. https://www.hathitrust.org/
7. http://creativecommons.org/licenses/
8. http://www.apache.org/licenses/LICENSE-2.0
9. http://www.gnu.org/licenses/gpl-3.0.en.html
10. https://developers.facebook.com/docs/graph-api
11. https://dev.twitter.com/overview/documentation
12. https://www.yelp.com/developers/documentation/v2/overview
13. http://orcid.org/
14. http://www-personal.umich.edu/~mejn/
15. http://mwnewman.people.si.umich.edu/
16. This data is often also used as additional features for machine learning tasks.
17. https://developers.google.com/freebase/?hl=en
18. The service was retired as of June 2015 and has been announced to be imported into Wikidata: https://www.wikidata.org/wiki/Wikidata:Main_Page
19. https://wordnet.princeton.edu/
20. http://globalwordnet.org/

References

Abello J, Broadwell P and Tangherlini TR (2012) Computational folkloristics. *Communications of the ACM* 55(7): 60–70.

Backstrom L, Boldi P, Rosa M, et al. (2012) Four degrees of separation. In: *Proceedings of the 4th ACM Web Science Conference (WebSci'12)*, Evanston, IL: ACM, pp. 33–42.

Cragin MH, Palmer CL, Carlson JR, et al. (2010) Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368(1926): 4023–4038.

De Choudhury M, Mason WA, Hofman JM, et al. (2010) Inferring relevant social networks from interpersonal communication. In: *Proceedings of the 19th International Conference on the World Wide Web (WWW'10)*, Raleigh, NC: ACM, pp. 301–310.

Diesner J (2013) From texts to networks: detecting and managing the impact of methodological choices for extracting network data from text data. *Künstliche Intelligenz/Artificial Intelligence* 27(1): 75–78.

Diesner J (2015) Words and networks: How reliable are network data constructed from text data? In: Bertino E and Matei S (eds) *Roles, Trust, and Reputation in Social Media Knowledge Markets*. Switzerland: Springer, pp. 81–89.

Diesner J and Evans CS (2015) Little bad concerns: Using sentiment analysis to assess structural balance in communication networks. In: *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2015)*, Paris, France.

Diesner J, Evans C and Kim J (2015) Impact of entity disambiguation errors on social network properties. In: *Paper presented at the International AAAI Conference on Web and Social Media (ICWSM)*, Oxford, UK: AAAI Press, pp. 81–90.

Fegley BD and Torvik VI (2013) Has large-scale named-entity network analysis been resting on a flawed assumption? *PLoS One* 8(7): e70299.

Fellbaum C (1998) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fellbaum C (2005) Wordnet and wordnets. In: Brown K (ed.) *Encyclopedia of Language and Linguistics*, 2nd ed. Oxford: Elsevier, pp. 665–670.

Gao J, Leetaru KH, Hu J, et al. (2013) Massive media event data analysis to assess world-wide political conflict and instability. *Lecture Notes in Computer Science: Vol 7812, Social Computing, Behavioral-Cultural Modeling and Prediction*. Berlin, Heidelberg: Springer, pp.284–292.

Graham J, Haidt J and Nosek BA (2009) Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96(5): 1029–1046.

Hammond J and Weidmann NB (2014) Using machine-coded event data for the micro-level study of political violence. *Research and Politics* 1(2): 1–8.

Howison J, Wiggins A and Crowston K (2011) Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems* 12(12): 767–797.

Iofciu T, Fankhauser P, Abel F, et al. (2011) Identifying users across social tagging systems. In: *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM-11)*, Barcelona, Spain: AAAI Press, pp. 1–4.

Kim J and Diesner J (2015a) Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. *Journal of the Association for Information Science and Technology* 66(9): 1–16.

Kim J and Diesner J (2015b) The effect of data pre-processing on understanding the evolution of collaboration networks. *Journal of Informetrics* 9(1): 226–236.

Kosinski M, Matz SC, Gosling SD, et al. (2015) Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist* 70(6): 543–556.

Lane J, Stodden V, Bender S, et al. (2014) *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. New York, NY: Cambridge University Press.

Lazer D, Pentland AS, Adamic L, et al. (2009) Life in the network: The coming age of computational social science. *Science* 323(5915): 721–723.

Moore R, Baru C, Rajasekar A, et al. (2000) Collection-based persistent digital archives-Part I. *D-Lib Magazine* 6(3): 1–16.

Ruths D and Pfeffer J (2014) Social media for large studies of behavior. *Science* 346(6213): 1063–1064.

Schrodt P, Gerner D and Yilmaz Ö (2004) Using event data to monitor contemporary conflict in the Israel-Palestine Dyad. In: *Paper presented at the International Studies Association*, Montreal, Quebec, Canada, pp. 1–31.

Schrodt P, Yilmaz Ö, Gerner D, et al. (2008) Coding sub-state actors using the CAMEO (Conflict and Mediation Event Observations) actor coding framework. In: *Paper presented at the Annual Meeting of the International Studies Association*, San Francisco, CA, pp. 1–39.

Talley EM, Newman D, Mimno D, et al. (2011) Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods* 8(6): 443–444.

Tausczik YR and Pennebaker JW (2010) The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1): 24–54.

Travers J and Milgram S (1969) An experimental study of the small world problem. *Sociometry* 32: 425–443.

Underwood T, Black ML, Auvil L, et al. (2013) Mapping mutable genres in structurally complex volumes. In: *Proceedings of the IEEE International Conference on Big Data (IEEE Big Data 2013)*, Santa Clara, CA: IEEE, pp. 95–103.

Wang F-Y, Carley KM, Zeng D, et al. (2007) Social computing: From social informatics to social intelligence. *Intelligent Systems, IEEE* 22(2): 79–83.

Wiebe J, Wilson T and Cardie C (2005) Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39(2): 165–210.

Wilson T, Wiebe J and Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, BC, Canada: ACL, pp. 347–354.

Zafarani R and Liu H (2013) Connecting users across social media sites: a behavioral-modeling approach. In: *Proceedings of the 19th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (KDD)*, Chicago, IL: ACM, pp. 41–49.

Zhao D and Strotmann A (2011) Counting first, last, or all authors in citation analysis: A comprehensive comparison in the highly collaborative stem cell research field. *Journal of the American Society for Information Science and Technology* 62(4): 654–676.

This article is part of a special theme on *Colloquium: Assumptions of Sociality*. To see a full list of all articles in this special theme, please click here: http://bds.sagepub.com/content/colloquium-assumptions-sociality.