Journal of Children and Media

ISSN: 1748-2798 (Print) 1748-2801 (Online) Journal homepage: http://www.tandfonline.com/loi/rchm20

Computational CAM: studying children and media in the age of big data

Brooke Foucault Welles

To cite this article: Brooke Foucault Welles (2016) Computational CAM: studying children and media in the age of big data, Journal of Children and Media, 10:1, 72-80, DOI: 10.1080/17482798.2015.1123165

To link to this article: http://dx.doi.org/10.1080/17482798.2015.1123165

JOURNAL OF CHILDREN AND MEDIA, 2016
VOL. 10, NO. 1, 72–80
http://dx.doi.org/10.1080/17482798.2015.1123165

Routledge
Taylor & Francis Group

Computational CAM: studying children and media in the age of big data

Brooke Foucault Welles

Department of Communication Studies, Northeastern University, Boston, MA, USA

ABSTRACT

New media technologies enable the pervasive and unobtrusive tracking of children's media use and interaction behaviors, and a host of new opportunities that researchers are only beginning to embrace. The possibilities for unlocking longstanding questions about children and media are thrilling, but the path forward is complicated by questions about epistemology, data, ethics and training. In this essay, I outline the promise and peril of computational social science for children and media research, drawing on advances from related fields to offer suggestions for researchers hoping to make use of big data and computational analytics.

ARTICLE HISTORY

Received 27 February 2015
Revised 8 July 2015
Accepted 27 July 2015

KEYWORDS

Big data; children and media; computational social science; new media; research agenda; research ethics

"We live life in the network," note Lazer and colleagues (Lazer et al., 2009) in their call for researchers to embrace *computational social science*, a new mode of social scientific inquiry involving the use of big data and computational analytics to examine human behavior. They describe a typical day for many affluent US Americans—a day characterized by interactions with and through technology—each of which leaves behind "digital breadcrumbs," or electronic records of individual, interpersonal, and group behavior. Computational social science is an interdisciplinary research approach that applies advanced computing methods to study these "digital breadcrumbs" as a means to understand human behavior (Cioffi-Revilla, Cioffi-Revilla, 2010). Related to studies of new media and computer-mediated communication more generally, computational social science is a distinct sub-field which relies on big data and computationally intensive research methods to analyze and model human social systems. Computational social science methods, such as automated text analysis, network analysis, and computational modeling, allow researchers to examine these social patterns over time, at massive scale, and across analytic levels in ways that were difficult or impossible in the recent past.

The methods of computational social science have already revolutionized the study of group dynamics, social influence, and political processes, among others (Aral & Walker, 2012; Bond et al., 2012; Zhu, Huang, & Contractor, 2013). As we look to the future, computational social science has the potential to revolutionize children and media research as well. The ecological and developmental theories and of children and media research are especially

CONTACT Brooke Foucault Welles b.welles@neu.edu

RCHM 1123165 9 January 2016 Revision CE: PJ QA: PM Coll:XX QC:XX

well matched to the analytic capabilities of computational social science, and appropriate big data are already abundant as children increasingly interact with and through technology, all the while, leaving behind data for analysis (Grimes & Fields, 2012).

Computational social science has the potential to unlock longstanding questions about children's interactions with and through media, including questions about the cumulative, longitudinal and widespread effects of children's media interaction. However, the path forward is complicated by questions about epistemology, data, ethics, and training. In this essay, I outline the potential and peril of computational social science for CAM research, with the goal motivating a conversation about when and how computational social science might be usefully applied to the study of children and media.

Big data, little kids

As recently as 15 years ago, computational research on children and media would have been impossible, because we lacked appropriate data. However, the technologies of the 21st century have shifted much of children's media consumption and production to mobile and/or online platforms, resulting in a surge of big data generated by and/or about children. Recent reports estimate that 95% of US adolescents use the internet, 81% are active on social media, and 78% have their own mobile phone (Madden et al., 2013; Madden, Lenhart, Duggan, Cortesi, & Gasser, 2013). Even very young children in the US and UK use new media technologies, with an estimated 72% of children aged 0–8 using mobile technologies to consume or produce content (Holloway, Green, & Livingstone, 2013; Rideout, 2013). These new media technologies, along with parallel advances in inexpensive recording technologies, enable a broad range of educational, social, and entertainment possibilities, and they also allow for the pervasive and unobtrusive tracking of children's educational, social, and entertainment behaviors—the so-called "digital breadcrumbs" that enable computational social science.

These data hold tremendous potential for CAM researchers, both as a complement to existing methods (survey, interview, etc.) and as pathway to ask new questions about children's media use. Although computational research is increasingly being incorporated in adjacent fields such as learning sciences (Romero & Ventura, 2007), the potential for computational CAM research remains largely untapped. The ways that computational social science could complement and extend CAM research are too numerous to elaborate in detail here, but consider the following illustrative examples:

Expanded data volume

One of the core strengths of computational social science is the ability to analyze very large data sets. Because digital data[1] can be interpreted by machines with minimal (if any) human coding, it is not uncommon for computational social science research to consider millions or even billions of data records in a single study. This presents some clear advantages in terms of scope, including the ability to collect and analyze near-comprehensive records of behaviors of interest. This possibility could enhance a number of CAM research areas including, for example, research on children's language acquisition. Using semi-automated text analysis algorithms applied and home videos of a child's first three years of life, Vosoughi and Roy (2012) have begun to make inroads in understanding how child—caregiver interactions influence the emergence of infant speech (see also Suskind, 2012). Semi-automated text analysis dramatically speeds the rate of audio transcription and analysis by automating much of the work, flagging only a small portion of difficult or unusual audio for human review and/or automatically extracting meaningful portions of text for further analysis (Roy, Vosoughi, & Roy 2010; Vosoughi & Roy, 2012). This means that in the time it used to take researchers to transcribe and annotate hours or days of material, researchers could feasibly transcribe and annotate weeks or months of recordings instead. Combined with the availability of inexpensive audio recording technologies, one could imagine expanding language acquisition research to include interactions with media such as television, radio and video-mediated calls (Skype, Facetime, etc.), thereby extending existing research on media and language acquisition (i.e., Krcmar, Grela, & Lin, 2007; Linebarger & Walker, 2005) to include near-comprehensive records of language exposure in the first year(s) of life. Such records could help overcome biases in recall and self-report on children's media exposure, and also complement qualitative research on language acquisition by identifying important patterns of media use for further in-depth research (Menchen-Trevino, 2013). Similar techniques could likewise be applied to any number of research areas involving the large-scale analysis of text. Indeed, automated text analysis has already been applied to study children's online communication (Velasquez et al., 2014), adolescent cyberbullying (Dinakar, Jones, Havasi, Lieberman, & Picard, 2012), and students' engagement in online earning environments (He, 2013), to name just a few.

Interdependent systems

In addition to handling large volumes of data, computational social science also has well-developed methods for analyzing large, interdependent systems. For example, researchers have already begun to apply a computational technique called "complex systems modeling" to understand the interdependent factors that influence educational outcomes (Maroulis et al., 2010). With a theoretical orientation evocative of Ecological Systems Theory (Bronfenbrenner, 1992), complex systems modeling allows researchers to examine how a number of interdependent micro-and meso-processes intersect to produce macro-level outcomes, all without specifying the nature of those interactions *a priori.* Instead, researchers input a variety of data that they suspect might be related to a particular outcome and computer algorithms test which combinations of data best predict the outcome(s) of interest (see Jacobson & Wilensky, 2006). One could apply this technique to any number of complex systems involving children and media, from the relationship between media consumption and aggressive behavior, to the influence of media on health, to and the effects of mediated communication on learning, and beyond. Moreover, because complex systems modeling integrates a number of different data sources, this technique is particularly complementary to more traditional methods, including survey and observational research, whose results can be used as variables in complex systems models.

Examining dynamic processes

A final example of where computational social science might be meaningfully integrated with CAM research is the examination of dynamic and/or emergent processes. Computational social science techniques increasingly focus on identifying dynamic processes that generate particular outcomes, rather than static inputs. For example, a growing body of research in network science examines how media consumption and interpersonal interaction combine to shape political attitudes and behaviors (Bond et al., 2012; Ksiazek, 2011). Rather than examining the net effect of various factors on political identity, techniques such as event-based network analysis can be applied to records of online behavior to model how each individual engagement with political news, advertising, and conversation influences all subsequent engagements with those materials (see Foucault Welles, Vashevko, Bennett, & Contractor, 2014). That is, event-based network models can identify dynamic changes in attitudes and behaviors over time. Of clear importance to many developmental processes, it is easy to imagine how such research could be applied to examine how children's political, social, educational, and entertainment preferences and capacities change over time. Moreover, with sufficiently diverse data sets, we may be able to use event-based network analysis and other dynamic computational techniques to identify when and how children's preferences and behaviors diverge along differences in gender, race, class, ability, or other dimensions of interest.

Challenges for computational CAM

The examples highlighted above are not meant to be comprehensive, rather provocative of near-term possibilities enabled by computational social science. The depth, breadth, and scale of research enabled by these techniques is thrilling, however this potential could be significantly encumbered by challenges in epistemology, data, ethics, and training. Such challenges are thoughtfully documented elsewhere (boyd & Crawford, 2012; Lazer et al., 2009; Vayena, Salathé, Madoff, Brownstein, & Bourne, 2015), but issues that may be particularly acute for research involving children warrant more attention here.

Epistemology

Advances in computational analytics enable interesting possibilities for research that extend and, in some cases, defy traditional modes of CAM theory building. In addition to more obvious changes in the scope and scale of data that can be included in analyses, large digital data sets also limit the need for statistical inference. It is difficult to imagine what it would mean to work with census data, and how existing CAM theories may be challenged as gaps in knowledge are filled. Take, for example, theories of media effects and aggressive behavior. Currently, the literature is broadly divided into studies of short-term effects observed in experiments, and long-term effects observed via panel studies and/or reflection surveys. Digital data have the potential to fill the space in between, documenting not only proximate media consumption and distal behavior, but also all the media consumed and many of the aggressive outbursts in between. Moreover, analytic techniques such as complex systems modeling and event-based network modeling, described above, will allow researchers to investigate relationships between media and aggression in such data sets without making statistical inferences or even specifying hypotheses in advance. This could leave CAM researchers in the unfamiliar position of knowing that a certain relationship between media and aggression exists, without a clear idea of what caused the effects they are observing. Such a predicament defies our conventional ways of understanding the children and media, although it not uncommon in other fields such as physics and genetics. However, sacrificing innovation for the sake of epistemological tradition hardly seems like the right answer here;

RCHM 1123165
9 January 2016
Revision
CE: PJ QA: PM
Coll:XX QC:XX

instead it will be important to carve out space in CAM conferences, workshops, seminars, and journals to share research results that apply computational techniques to advance theory in ways that were not previously possible.

Data and ethics

Obtaining and ethically managing data is arguably the largest challenge for computational CAM research. Typically, computational social science relies on "found" data sets, often including online data that were not collected for research purposes (Lazer et al., 2009). The Children's Online Privacy Protection Act (COPPA) limits the collection of such data about children less than 13 years old, requiring verifiable parental consent prior to collecting personally identifying information online. As a result, despite a growing number of educational and entertainment services targeted at children, there is scant evidence of the availability—or even existence—of data collected from these sites (Grimes & Fields, 2012). Although much of the computational social science research described above can be conducted in aggregate, and without access personally identifying information about individual children (and would therefore not be in violation of COPPA regulations), it is rare to see research involving direct access to records of children's online behavior, and rarer still for such data to be freely available[2]. The (in)availability of data, while in the best interest of individual children, presents a number of problems for social scientific inquiry and could limit the potential to conduct computational CAM research to private companies working with proprietary data sets that can neither be examined nor used for replication (Lazer et al., 2009).

To avoid this kind of cloistering of scientific potential, while still protecting the interests of young children, it is critical that the CAM community actively pursue opportunities for informed data sharing and organizational-academic partnerships. One pathway forward may be to leverage online data donation portals such as *volunteerscience.com* that allow individual users (and their parents) to donate data for scientific research, with full informed consent. Compliant with COPPA, and ethically consistent with CAM research conducted in laboratory settings, this option would allow researchers to develop a "virtual laboratory" of computational data generated by and about children.

However, relying on data donations alone subverts some of key benefits of computational social science, reproducing problems with biased samples and missing data. As a supplement, it will be important to establish research partnerships with organizations that have already developed techniques to collect online data about children. Fortunately, unlike in many other fields, there are already outstanding models of successful partnerships around research involving children and media in the non-profit sector. For example, the Sesame Workshop, Joan Ganz Cooney Center, and Corporation for Public Broadcasting all support research fellowships, internships, post-doctoral students, and/or university collaborations that include access to children's television and digital media content. These collaborations may provide a starting point for access to online data more generally, or, at minimum, serve as examples for how these partnerships could work.

It will, of course, be important to work out issues of privacy, particularly with regard to data that were not collected expressly for research purposes. Even in cases where such data were collected in compliance with COPPA regulations, and especially when COPPA does not apply (for children 13+), it is incumbent upon CAM researchers to be vigilant about issues of informed consent, data security, and participants' privacy. Notably, online data are typically not organized in a way that maximally protects participants, nor can they easily be shared without risking privacy breaches (Narayanan & Shmatikov, 2010). As evidenced by public relations debacles for companies like AOL and Facebook (Barbaro & Zeller, 2006; Goel, 2014), using data for research purposes may introduce real or perceived privacy violations, especially for users who were not aware they were participating in scientific research. To that end, it will be important for the CAM community to establish guidelines for ethical collection, analysis, sharing, and reporting on research involving electronic data collected from and about children.

A final complication with data involves participants' access to the devices generating data of interest. Although children are increasingly using the Internet, social media, and mobile devices, access to these technologies remains far from universal, and limits on access and use disproportionately affect certain populations, including children with disabilities and low-income youth of color (Hargittai, 2010). These children are already underrepresented in social science research more generally (Henrich, Heine, & Norenzayan, 2010), and their limited data may be the most likely to be trimmed from large data sets if we do not make concerted efforts not to do so (Welles, 2014). These biases may be at least partially remedied by actively seeking out data generated by underrepresented children, for example, by studying technology engagement in urban school districts (Gomez & Pinkard, 2014), or by specifically cultivating data sharing agreements with companies that design technologies for youth with disabilities.

Training

Beyond issues of theory, data, and ethics, there are very important considerations relating to the basic skills and appropriate training of graduate students and faculty interested in computational CAM research. Currently, there are few off-the-shelf tools that allow researchers to easily download, organize, and/or analyze computational data. Although it is possible this may change in the future, social scientists have been bemoaning the divides created by the dearth of tools for computational research for years, with little noticeable progress (boyd & Crawford, 2012; Lazer et al., 2009). So, for the time being, it seems safe to assume that a minimum entry skill set for computational CAM research will include basic coding and scripting skills, along with an understanding of databases and methods to query them. Closely related fields, such as Learning Sciences, have integrated technical training into many graduate programs—a strategy that is plausible for CAM in the long term as well. Alternatively, we might adopt Lazer and colleagues' (2009) suggestion to rely on collaborative teams of "computationally literate social scientists and socially literate computer scientists," (p. 722). Such a strategy may be more tractable than revamping PhD curriculum in the short term, but will need to be embraced by the CAM community, with particular recognition for the difficulty and value of interdisciplinary collaboration in publishing, hiring, promotion, and tenure decisions.

Conclusion

The possibilities for using digital data and computational analytics to study children and media are exhilarating. Although the path forward is complicated by questions about epistemology, training, data, and ethics, these issues are not insurmountable, and the rewards for negotiating these issues could be great—for CAM and computational social science alike. Although this essay focuses on how CAM might benefit from computational research, computational research stands to benefit from CAM as well. Notably, the theories of CAM are already ecological and/or developmental; adapting computational methods to accommodate these theories represents a unique opportunity for innovation, one that is a natural fit for both sides. Similarly, guidelines established to protect children's data privacy will serve all computational researchers well, even those working with adults' data. Finally, integrating CAM—a field with a historically high proportion of women researchers—into computational social science early in its emergence as an interdisciplinary field may help computational social science avoid the persistent and pervasive gender disparities found elsewhere in computing.

In sum, my hope is that this essay will prompt conversation about the opportunities and risks of computational CAM research, neither of which have been fully explored here. Among the opportunities, I have only scratched the surface of possible paths new research could take, and ignored almost entirely potential for triangulation, validation and replication enabled by introducing a new suite of computational methods into CAM research. Similarly, although I have attempted to outline the most serious risks for emerging computational CAM, I have not discussed many simple complications, such as the logistics of setting up and maintaining computing infrastructure, or matching data, methods and research questions about children of different ages, or more difficult issues, such as open data and the ethics of sharing data within research communities. In the next ten years, I hope to see conferences, workshops, and special issues dedicated to hashing out the opportunities, risks and responsibilities of computational CAM research, so that we might catalyze a new wave of inquiry that leverages contemporary data and analytics to answer longstanding questions in the field.

Notes

1. Digital data are represented as discrete, binary digits that can be read by computers. In contrast, analog data include continuous waves that fluctuate within an infinite range. Much of the data considered here are "born digital," or generated by computers in digital form. For practical purposes, even traditionally analog data (such as audio recordings) are now converted to digital form for storage purposes (e.g., converting LP audio to MP3), so even data that were not "born digital" can often be accessed that way.
2. Notable exceptions include Kafai (2010) and Resnick et al. (2009); both excellent examples of computational analyses conducted without personally identifying information.

Acknowledgments

I would like to thank David Lazer, Benjamin Mako Hill, and two blind reviewers for their helpful suggestions as I prepared this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was financially supported in part by the interdisciplinary Faculty Research and Creative Activity Grant from Northeastern University's College of Arts, Media and Design.

Notes on contributor

Brooke Foucault Welles is an assistant professor in the department of Communication Studies and a faculty affiliate of the Network Science Institute and NU Lab for Texts, Maps and Networks at Northeastern University. Foucault Welles studies how social networks shape and constrain human behavior, with a particular emphasis on how young people use social networks in the pursuit personal, group, and social goals.

References

Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science, 337*, 337–341. doi:10.1126/science.1215842

Barbaro, M., & Zeller, T., Jr. (2006, August 9). A face is exposed for AOL searcher No. 4417749. *The New York Times*. Retrieved December 9, 2015. http://www.nytimes.com/2006/08/09/technology/09aol.html?_r=2&pagewanted=all

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*, 295–298.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*, 662–679.

Bronfenbrenner, U. (1992). Ecological systems theory. In R. Vasta (Ed.), *Six theories of child development: Revised formulations and current issues* (pp. 187–249). London: Jessica Kingsley Publishers.

Cioffi-Revilla, C. (2010). Computational social science. *Wiley interdisciplinary reviews: Computational statistics, 2*, 259–271.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS), 2*, article 18.

Foucault Welles, B., Vashevko, A., Bennett, N., & Contractor, N. (2014). Dynamic models of communication in an online friendship network. *Communication Methods and Measures, 8*, 223–243.

Goel, V. (2014, June 29). Facebook tinkers with users' emotions in news feed experiment, stirring outcry. *The New York Times*. Retrieved December 9, 2015. http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html

Gomez, K., & Pinkard, N. (2014). *The digital youth network: Cultivating digital media citizenship in urban communities*. Cambridge, MA: MIT Press.

Grimes, S., & Fields, D. (2012). *Kids online: A new research agenda for understanding social networking forums*. New York: The Joan Ganz Cooney Center at Sesame Workshop.

Hargittai, E. (2010). Digital na(t)ives? Variation in internet skills and uses among members of the "Net Generation". *Sociological Inquiry, 80*, 92–113.

He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior, 29*, 90–102.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–83.

Holloway, D., Green, L., & Livingstone, S. (2013). *Zero to eight: Young children and their internet use*. London: EU Kids Online, LSE.

Jacobson, M. J., & Wilensky, U. (2006). Complex systems in education: Scientific and educational importance and implications for the learning sciences. *Journal of the Learning Sciences, 15*, 11–34.

Kafai, Y. B. (2010). World of Whyville an introduction to tween virtual life. *Games and Culture, 5*, 3–22.

Krcmar, M., Grela, B., & Lin, K. (2007). Can toddlers learn vocabulary from television? An experimental approach. *Media Psychology, 10*, 41–63.

Ksiazek, T. B. (2011). A network analytic approach to understanding cross-platform audience behavior. *Journal of Media Economics, 24*, 237–251. doi:10.1080/08997764.2011.626985

Lazer, D., Pentland, A. S., Adamic, L. A., Aral, S., Barabasi, A. L., Brewer, D., … van Alstyne, M. (2009). Computational social science. *Science, 323*, 721–723.

Linebarger, D. L., & Walker, D. (2005). Infants' and toddlers' television viewing and language outcomes. *American Behavioral Scientist, 48*, 624–645.

Madden, M., Lenhart, A., Cortesi, S., Gasser, U., Duggan, M., Smith, A., & Beaton, M. (2013). *Teens, social media and privacy*. Washington, DC: Pew Internet and American Life Project.

Madden, M., Lenhart, A., Duggan, M., Cortesi, S., & Gasser, U. (2013). *Teens and technology 2013*. Washington, DC: Pew Internet & American Life Project.

Maroulis, S., Guimera, R., Petry, H., Stringer, M., Gomez, L., Amaral, L., & Wilensky, U. (2010). Complex systems view of educational policy research. *Science, 330*, 38–39.

Menchen-Trevino, E. (2013). Collecting vertical trace data: Big possibilities and big challenges for multi-method research. *Policy & Internet, 5*, 328–339.

Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of personally identifiable information. *Communications of the ACM, 53*, 24–26.

Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., … Silverman, B. (2009). Scratch: Programming for all. *Communications of the ACM, 52*, 60–67.

Rideout, V. J. (2013). *Zero to eight: Children's media use in America 2013 Program for the Study of Children and Media*. San Francisco, CA: Common Sense Media.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications, 33*, 135–146.

Roy, B. C., Vosoughi, S., & Roy, D. K (2010). *Automatic estimation of transcription accuracy and difficulty*. Paper presented at the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010), Makuhari, Chiba, Japan

Suskind, D. (2012). *A randomized controlled pilot of a novel parent-directed language intervention in children of low SES: The thirty million words project*. Paper presented at the 2012 AAP National Conference and Exhibition. New Orleans, LA.

Vayena, E., Salathé, M., Madoff, L. C., Brownstein, J. S., & Bourne, P. E. (2015). Ethical challenges of big data in public health. *PLOS Computational Biology, 11*, e1003904.

Velasquez, N. F., Fields, D. A., Olsen, D., Martin, T., Shepherd, M. C., Strommer, A., & Kafai, Y. B. (2014). *Novice programmers talking about projects: What automated text analysis reveals about online Scratch users' comments*. Paper presented at the 47th Hawaii International Conference on System Sciences (HICSS), Waikoloa, HI, USA.

Vosoughi, S., & Roy, D. K. (2012). *An automatic child-directed speech detector for the study of child language development*. Paper presented at the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012). Portland, OR.

Welles, B. F. (2014). On minorities and outliers: The case for making big data small. *Big Data & Society, 1*(1), 1–2.

Zhu, M., Huang, Y., & Contractor, N. S. (2013). Motivations for self-assembling into project teams. *Social Networks, 35*, 251–264.