Xu *EPJ Data Science* (2023) 12:61
https://doi.org/10.1140/epjds/s13688-023-00436-z

EPJ Data Science
a SpringerOpen Journal

COMMENTARY Open Access

Studying social networks in the age of computational social science

Xinwei Xu[1*]

[*]Correspondence:
xinwei.xu@gess.ethz.ch
[1]Department of Humanities and Social Sciences, ETH Zurich, Weinbergstrasse 109, 8092 Zurich, Switzerland

Abstract

Social and behavioral sciences now stand at a critical juncture. The emergence of Computational Social Science has significantly changed how social networks are studied. In his keynote at IC2S2 2021, Lehmann presented a series of research based on the Copenhagen Network Study and pointed out an important insight that has mostly gone unnoticed for many network science practitioners: the data generation process — in particular, how data is aggregated over time and the medium through which social interactions occur — could shape the structure of networks that researchers observe. Situating the keynote in the broader field of CSS, this commentary expands on its relevance for the shared challenges and ongoing development of CSS.

Keywords: Computational social science (CSS); Dynamic networks; Data aggregation

1 Introduction

Few would disagree that social and behavioral sciences stand at a critical juncture of development. More than ever before, researchers have access to a vast wealth of information and digital records that could potentially be used as "data"; meanwhile, the advancement of computational tools allows researchers to process and analyze those data with tremendous analytical power that was unthinkable just a few decades ago. Computational Social Science (CSS) emerges as a timely child of all these advents [1–7]. CSS involves a multidisciplinary effort with "the development and application of computational methods to complex, typically large-scale, human (sometimes simulated) behavioral data" [1]. Unlike traditional datasets curated and collected by social scientists, these new data are often mined from the "digital footprint" [2] of massive online behaviors and interactions, digitized archives of administrative records [4], and increasingly, digital traces of off-line interactions and activities via sensor technologies [8–10].

At the core of the rising waves of CSS is the idea of networks, both in its theoretical and methodological form. Perhaps no other contemporary scientific idea can spawn such wide-ranging disciplinary interests. Textbooks and handbooks on Network Theories and Methods have been written by communication scholars, computer scientists, economists, physicists, political scientists, psychologists, sociologists, and statisticians (see, for example, [11–17]), and the list can go on. In the most general term, a network consists of "nodes" and “edges” linking those nodes [15]. Like any other tool of scientific abstraction, it is a simplified representation distilled from complex social phenomena.

The study of social networks — networks where nodes represent people and edges represent some form of connection between them — can be traced back to the 1930s [15, 18]. Ever since, what constitutes a social network and how it is measured and represented have significantly changed as different generations of researchers have had very different data at their disposal. The spectrum of social network data has evolved from ethnographic observations conducted by early pioneers in network research to qualitative interviews and quantitative surveys [18], and more recently, encompassing digital records. As the boundaries of online- and offline-world become increasingly entangled [19], studying social networks implies a multifaceted endeavor. The Copenhagen Network Study ([20]; CNS) highlighted in Lehmann’s keynote speech is one among such recent data collection efforts [21, 22]. Presenting a series of research based on the CNS, Lehmann [23] pointed out a critical insight that has mostly gone unnoticed for many network science practitioners: the data generation process — in particular, how data is aggregated over time and the medium through which social interactions occur — could shape the structure of networks that researchers observe.

This commentary first gives a summary of the keynote speech. Next, situating the keynote in the broader field of CSS, I show the different approaches of the two main branches of CSS — computational sciences versus social sciences. Then I argue that the insight from the keynote points to some of the shared challenges on both sides. Most importantly, there exists a trade-off between prioritizing the richness of data and extracting scientific meanings out of such data. The final section suggests two ongoing trends in the field of CSS that could help address those shared challenges — data triangulation and the integration of data-driven and theory-driven research.

2 Summary of the keynote

Throughout the keynote, Lehmann’s main message is that the way data is generated, aggregated, and represented shapes the structure of networks. Previous research on identifying clusters within networks, called “community detection”, has mostly focused on partitioning nodes into communities given the patterns of connections between them. However, as nodes tend to belong to multiple communities, the pervasive overlap of communities leads to networks that are locally dense and globally sparse. A solution to this problem is to focus on clustering links rather than nodes [24]. In a series of studies centered around the CNS [20] — a longitudinal high-resolution interaction network of over 700 undergraduate students based on bluetooth, phone-logs, and Facebook data, Lehmann extended this line of thought and suggested that the unique properties of “links”, such as its temporal resolution and the interaction medium, could have important implications for how we ascertain the network structure.

First, Lehmann and colleagues found that looking at data at different time scales can reveal different groupings within the network. This is because the process of aggregating data over a longer period can obscure the unique characteristics of networks with distinct fundamental structures [25]. For instance, in one study, [26] argued that the complexity of social networks can be simplified by breaking down the data into shorter time periods, or micro-episodes of “gatherings.” They showed that networks that appear dense and complex on a daily basis could be decomposed into sparser networks when analyzed at hourly or 5-minute intervals. These structures, which are formed by repeated gatherings over time, contain stable "cores" and can accurately predict individuals' geosocial behaviors based on human mobility models [27].

Lehmann's second key observation was that the medium through which social interactions occur (such as Bluetooth sensors, phone logs, or social media) also influences the structure of networks. While networks formed through different communication channels may have similar structural properties when analyzed over a long period of time, the microscopic structures of these networks can be very different. For example, phone networks are made up of pairs of individuals (dyads) because phone calls typically only involve two people, which limits the ability for triadic closure (three people connecting synchronously). In contrast, the basic units of Facebook networks consist of one-to-many interactions that may not be synchronous. Lehmann proposed two dimensions for classifying these microscopic structures: whether the communication is synchronous or asynchronous, and whether it is one-to-one, one-to-many, or many-to-many [25]. Each type of communication creates a unique "network footprint," or "dynamical class" of networks, and the corresponding data collection process produces different results at different time scales.

Altogether, these two insights challenge some of the existing research practices in network science. For example, many of the null models that researchers rely on in generating a reference distribution of graphs often result in microscopic configurations that are completely unrealistic (such as random-mixing graphs). In addition, comparing counts of network motifs [28] is only meaningful when the data generating processes are comparable, or, in [23]'s own term, belonging to the same "dynamical class" of networks. More generally, it also raises the question to what extent we could generalize the patterns and properties observed across networks with very different underlying data generation processes. Beyond these immediate implications for network science, I expand on the relevance of Lehmann's keynote to the broader field of CSS.

3 Relevance to CSS

Although what CSS encapsulates has shifted over the past decades [4], the evolving field of CSS is centered around a shared interest in the ever-expanding reservoirs of digital data and a shared goal of detecting and making sense of patterns of complex human behaviors revealed by those data. As the field of CSS emerges, there exists a natural division of labor between the social science side and the computational side. Social sciences provide questions, while computational sciences enable social scientists to answer those questions with new tools and new perspectives, consisting of a "trading zone" between different disciplines [3].

On the one hand, the Copenhagen Network Study, alongside the series of research based on it, illustrates the different disciplinary approaches of doing science under the loose umbrella of CSS — particularly, the difference between prediction-oriented and explanation-oriented research. On the other hand, the insights raised by Lehmann in the keynote point to some common challenges shared by both the computational side and the social science side of the CSS. It is not only the similar interests in digital data that these two sides share but also conundrums.

3.1 Different disciplinary approaches

In the most simplistic division, the computational side of the CSS is mostly driven by data and prioritizes *prediction-oriented* research, while the social science side of the CSS mostly orients towards theory and prioritizes *explanation-oriented* research [6]. This difference can be seen from a highly simplified comparison of how each side deals with high-resolution temporal data on communication events.

In the study highlighted in the keynote [26], the goal of the researchers is to detect meaningful patterns from high-frequency data on communication networks. When the researchers de-aggregate the temporal dimension and slice the network with shorter time intervals, the network structure becomes apparent and gets simplified. The researchers further find that the stable "cores" that emerged from those repeated gatherings over time can precisely predict the geosocial behaviors of individuals. In this case, the network itself consists of "links" that represent instances of communication events rather than measurement for a specific theoretical construct. Thus reducing the data complexity happens at the stage of data analysis, where the unit of observation (a communication event) overlaps with the unit of analysis. Eventually, the researchers are interested in how well those structures can predict individuals' behaviors within a certain time span.

In an early sociological study [29], the researchers are confronted with a similar type of data albeit less multifaceted than the CNS — minute-to-minute email exchanges at an American university. The goal of the researchers is to explain how "ties" form over time. Thus the first item on the researchers' to-do list is to construct an empirically reliable and theoretically valid measurement for what constitutes a "tie". The formation and the dissolution of a tie are functions of the intensity of the communication events. In this case, reducing the data complexity happens at the measurement construction stage, prior to any kind of analysis and modeling. The unit of analysis (a tie) is an aggregated abstraction from the unit of observation (a communication event). Eventually, the researchers arrive at a theoretical explanation for how ties form and why they manifest into the structural forms they do.

Although the prediction-oriented and the explanation-oriented approach are not necessarily mutually exclusive, each involves very different analytical choices regarding measurement construction. For a task of predicting individuals' mobility patterns within a relatively short time span, it lends researchers more predictive power when construing "links" on a minute-to-minute or hourly basis. In contrast, for a task of explaining how individuals become friends with a specific type of people (say, people with similar income, similar professions, living in similar neighborhoods, affiliated with the same sport club), it often requires a theoretical construct (a tie) that goes beyond the unit of observation (a communication event).

Two lessons become apparent from this simple exercise. First, having rich data could yield novel insights, yet having rich data cannot substitute the process of understanding the "why" behind the patterns that emerged from the data. Sometimes domain knowledge is also required. This also distinguishes CSS from pure prediction tasks that commonly populate the industry of data science. Second, the theories we enlist to explain the "why" can have unspoken limits and assumptions — such as their temporal scope. Sometimes those assumptions are not directly transferable between different types of data and different contexts.

3.2 Shared challenges

The main message delivered in the keynote — that the data generation process shapes patterns that emerge from the data — is not unique to computational scientists. We see similar lines of inquiry in the development of traditional social science as well. Take, for instance, the well-established survey method that has dominated social sciences for the past few decades. The line of literature on survey methodology [30] and the psychology behind survey response [31] shows the importance of how the data generating process should be considered when researchers design and field a survey — such that the potential biases induced by certain data generation practices could be minimized.

In fact, when we look at the development of social sciences over the past century, the emergence of a dominating research paradigm is invariably accompanied by the availability of new data sources and research tools brought about by the advent of new technologies. For instance, the prevalence of survey data is a result of the availability of phones and later computers, which largely facilitated the ways surveys are conducted and reduced the costs of large-scale surveys. Nowadays, with the possibility of online surveys and the boom of survey companies, it is not uncommon that researchers can tap into a large pool of potential survey respondents and experiment subjects just within a few clicks (e.g., [32, 33]). With those in mind, we see that the availability of new data and tools inevitably shapes the social reality within the grasp of the researchers. Thus it is almost unavoidable that the dominating ways of doing social sciences will be challenged by new methods and the new representation of social reality that they produce.

To this end, the issues raised in the keynote represent an instance where both sides (computational sciences and social sciences) have been grappling with a similar challenge: there exists a non-trivial trade-off between prioritizing the richness of data and extracting scientific meanings out of such data. More specifically, how can we effectively reduce large amounts of data to a more manageable size while still maintaining its richness and depth, and what patterns can we identify in the process? [3] Do those patterns represent new instances of existing theories, or do we need new theory-building efforts in order to understand those "new" phenomena?

Along this line, a key methodological challenge in network research involves deciding what constitutes a tie. As mentioned above, for computational sciences, the unit of observation (e.g., a communication event) and the unit of analysis (e.g., a network tie) are not clearly demarcated; for social sciences, however, the unit of analysis often requires certain processes of aggregation and abstraction from the unit of observation. For instance, in traditional survey data, the process of abstraction is based on the cognitive representation of the survey respondent (e.g., by answering "who is your best friend in the classroom"?). With passive behavioral data such as those extracted from phone logs, email exchanges, and social media, the task of abstraction falls onto the researcher — when and how do we determine what a meaningful tie is? Is such abstraction consistent with how the actors themselves would otherwise perceive it?

This is further complicated by the fact that the measurement itself can interfere with what we intend to measure [7]. As pointed out in Lehmann (2021)'s keynote, the microscopic structures of communication are highly dependent on the type of instrument we use and the design of the platforms or technologies that we rely on (e.g., sensors, phone logs, social media, etc.). Thus this requires researchers to be aware of the constraints and the unspoken assumptions that the data generating process could impose.

4 Future development of CSS

4.1 Data linking and triangulation

Going back to the shared challenges, we see that both traditional and new digital data have limits in measuring the theoretical constructs that are often at the core of social science theories. This calls for a need for data merging, cross-platform data linking (e.g., see Adam's keynote [34]), and data triangulation [7].

In the early heyday of "big data", digital traces — such as those scrapped from online shopping platforms and social media — are often passively collected as human behaviors unfold in real-time and thus are not particularly "designed" for research purposes [35]. Compared to traditional datasets that are often produced and curated under the supervision of the researchers — such as those from surveys and lab experiments, those "naturalistic" data are seen by social scienctists with an eye of suspicion [36, 37] due to issues such as data representation and confounding measurements. Against this background, the Copenhagen Network Study is unique in that it showcases a potential solution for some of those criticisms. It is among a few recent cases (e.g., see other similar data collection efforts in [21, 22]) that not only show that a multifaceted data collection strategy with a large population is feasible but also that such data can be specifically harnessed and curated for research purposes (although not without ethical concerns).

The need for data triangulation is also called for by the limits of existing social theory [7]. Having a more realistic grasp of social reality could not only help researchers interrogate the validity of their measurements but also inform theory development — particularly in locating theoretical blind spots. In Lehmann (2021)'s example, the idea that different temporal aggregation of communication events results in different network structures immediately brings the question of temporality to the front end. While longitudinal analyses of networks have become a routine part of the scholarly dialogue [38], the theory-building effort has focused chiefly on static network structures and the dominating network processes that lead to such structures. We lack a realistic theory of how networks evolve (and potentially at different time scales). For instance, most network studies focus on tie formation, yet very few look into tie decay and dissolution. Although we have well-established theories on network mechanisms such as homophily, reciprocity, and transitivity [39], there is a lack of attention on how those mechanisms could vary over time [40] and to what extent they could be generalized into digital networks [19].

4.2 Integrating data-driven and theory-driven research

As different disciplines have different approaches of doing science, data-driven research and theory-driven research have mostly developed along segregated lines. Most often, we see a one-sided import of computational methods into social sciences [2, 4, 41]. In contrast, Lehmann (2021)'s own trajectory of extracting theories ("dynamical class" and "fundamental structures") from empirical insights illustrates a possible reverse flow of knowledge. Those theory-building efforts are needed to synthesize and integrate pockets of empirical insights that emerged from ongoing streams of data-driven research. Additionally, the realization that his own classification of communication events overlaps with theoretical frameworks developed earlier by communication scholars [42] precisely shows a need for more two-sided exchange between theory-driven and data-driven research.

In sum, although the epistemic values of the social science side and the computational side differ [3], they could offer complementary perspectives, and the integration of both could yield fruitful advances in scientific insights [6]. As it is often the case that scientific papers conclude with limitations of the current study and suggest potential future directions, the existing limits of the data-driven approach and the existing limits of social theory present tremendous opportunities for interdisciplinary collaboration and, potentially, paving the way for the emergence of a new research paradigm. Thus as data-driven researchers become more theoretically-minded and theory-driven researchers become more data-enabled, there is hope that we see some converging points on a unified framework for research practices and methodologies. Yet, integrating different strands and possibly different paradigms of research requires "brokers" that could link different scholarly communities. So far, it still remains a challenge to facilitate such "bridging" ties and interdisciplinary dialogues, as the institutional incentives and the corresponding infrastructures are still wanting [5].

Acknowledgements
For their careful reading and helpful comments on previous drafts, I am grateful to Aliakbar Akbaritabar, Timon Elmer, Christoph Stadtfeld, and participants at the Human and Networks in a Digital Age Seminar at ETH Zurich. I would also like to thank Termeh Shafie and Christoph Stadtfeld for the opportunity to contribute to this special issue and Sune Lehmann for an inspiring keynote speech at IC2S2.

Funding
Not applicable.

Data availability
Not applicable.

Declarations

Competing interests
The author declares no competing interests.

Author contributions
XX wrote and revised the manuscript. The author read and approved the final manuscript.

Received: 8 January 2023 Accepted: 30 November 2023 Published online: 19 December 2023

References

1. Lazer D, Pentland A, Adamic L et al (2009) Computational social science. Science 323:721–723
2. Golder SA, Macy MW (2014) Digital footprints: opportunities and challenges for online social research. Annu Rev Sociol 40:129–152
3. McFarland DA, Lewis K, Goldberg A (2016) Sociology in the era of big data: the ascent of forensic social science. Am Sociol 47:12–35
4. Edelmann A, Wolff T, Montagne D, Bail CA (2020) Computational social science and sociology. Annu Rev Sociol 46:61–81
5. Lazer DM, Pentland A, Watts DJ et al (2020) Computational social science: obstacles and opportunities. Science 369:1060–1062
6. Hofman JM, Watts DJ, Athey S et al (2021) Integrating explanation and prediction in computational social science. Nature 595:181–188
7. Wagner C, Strohmaier M, Olteanu A et al (2021) Measuring algorithmically infused societies. Nature 595:197–204
8. Eagle N et al (2006) Reality mining: sensing complex social systems. Pers Ubiquitous Comput 10:255–268
9. Sapiezynski P, Stopczynski A, Wind DK et al (2017) Inferring person-to-person proximity using WiFi signals. In: Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, vol 1, pp 1–20
10. Elmer T, Chaitanya K, Purwar P, Stadtfeld C (2019) The validity of RFID badges measuring face-to-face interactions. Behav Res Methods 51:2120–2138
11. Wasserman S, Faust K et al (1994) Social network analysis: methods and applications
12. Monge PR, Contractor NS, Contractor PS et al (2003) Theories of communication networks. Oxford University Press, London
13. Easley D, Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, Cambridge
14. Jackson MO (2010) Social and economic networks. Princeton University Press, Princeton
15. Newman M (2018) Networks. Oxford University Press, London
16. Cranmer SJ, Desmarais BA, Morgan JW (2020) Inferential network analysis. Cambridge University Press, Cambridge
17. Light R, Moody J (2020) The Oxford handbook of social networks. Oxford University Press, London
18. Freeman L (2004) The development of social network analysis. Empirical Press, Vancouver

19. Lewis K (2021) Digital networks: elements of a theoretical framework. Soc Netw
20. Sapiezynski P, Stopczynski A, Lassen DD, Lehmann S (2019) Interaction data from the Copenhagen networks study. Sci Data 6:1–10
21. Wang C, Lizardo O, Hachen DS (2020) Neither influence nor selection: examining co-evolution of political orientation and social networks in the NetSense and NetHealth studies. PLoS ONE 15:e0233458
22. Vörös A, Boda Z, Elmer T et al (2021) The Swiss StudentLife study: investigating the emergence of an undergraduate community through dynamic, multidimensional social network data. Soc Netw 65:71–84
23. Lehmann S (2021) How the data generating process shapes dynamic networks. In: IC2S2 Keynotes
24. Ahn Y-Y, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. Nature 466:761–764
25. Lehmann S (2019) Fundamental structures in temporal communication networks. In: Temporal network theory. Springer, Berlin, pp 25–48
26. Sekara V, Stopczynski A, Lehmann S (2016) Fundamental structures of dynamic social networks. Proc Natl Acad Sci 113:9977–9982
27. Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of predictability in human mobility. Science 327:1018–1021
28. Milo R, Shen-Orr S, Itzkovitz S et al (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827
29. Kossinets G, Watts DJ (2009) Origins of homophily in an evolving social network. Am J Sociol 115:405–450
30. Groves RM, Fowler FJ Jr, Couper MP et al (2011) Survey methodology. Wiley, New York
31. Tourangeau R, Rips LJ, Rasinski K (2000). The psychology of survey response
32. 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. Am Psychol 70:543–556. https://doi.org/10.1037/a0039210
33. Coppock A, McClellan OA (2019) Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. Res Polit 6:2053168018822174
34. Adam S (2021) Pushing research on user-centric information exposure forward: bringing tracking, survey and automated text classification together. In: IC2S2 keynotes
35. Salganik MJ (2019) Bit by bit. Social research in the digital age. Princeton University Press, Princeton
36. Lewis K (2015) Three fallacies of digital footprints. Big Data Soc 2:2053951715602496
37. González-Bailón S, Wang N, Rivero A et al (2014) Assessing the bias in samples of large online networks. Soc Netw 38:16–27
38. Snijders TA, Doreian P (2010) Introduction to the special issue on network dynamics. Soc Netw 32:1–3
39. Rivera MT, Soderstrom SB, Uzzi B (2010) Dynamics of dyads in social networks: assortative, relational, and proximity mechanisms. Annu Rev Sociol 36:91–115
40. Lewis K, Kaufman J (2018) The conversion of cultural tastes into social network ties. Am J Sociol 123:1684–1742
41. Molina M, Garip F (2019) Machine learning for sociology. Annu Rev Sociol 45:27–45
42. Jensen KB, Helles R (2011) The Internet as a cultural forum: implications for research. New Media Soc 13:517–533

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.