Representative methods of computational socioeconomics

Tao Zhou ∗
CompleX Lab, University of Electronic Science and Technology of China, Chengdu 611731, People’s Republic of China
∗ Author to whom any correspondence should be addressed.
E-mail: zhutou@ustc.edu

Keywords: computational socioeconomics, big data, natural data, online experiments, survey expanding

Abstract

The increasing availability of data sources and analysis tools borrowed from computer science and physical science have sharply changed traditional methodologies of social sciences, leading to a new branch named computational socioeconomics, which studies various phenomena in socioeconomic development by using quantitative methods based on large-scale real-world data. Situated on recent publications, this perspective will introduce three representative methods: (i) natural data analyses, (ii) large-scale online experiments, and (iii) integration of big data and surveys. This perspective ends up with in-depth discussion on the limitations and challenges of the above-mentioned emerging methods.

1. Introduction

Social sciences study the social structure based on activities of and relations between human beings, encompassing a wide array of academic disciplines including sociology, economics, politics, linguistics, jurisprudence, and many other branches. In spite of the intrinsic difficulty of quantification, social sciences recently show a higher and higher level of quantification and become increasingly dependent on real-world data [1–5]. However, the traditional way to obtain real-world data has many limitations. For example, survey data from questionnaires usually contains insufficient samples for solid statistics and suffers from social desirability bias [6], while large-scale and precise censuses consume huge resources and lack timeliness.

Fortunately, the ongoing digital wave results in unprecedentedly rich data that records microscopic processes of social and economic development, including satellite remote sensing data, mobile phone data, social media data, and so on. These data are usually of large sizes, almost in real time and with high resolution, which can reduce the small-sample bias and the invisible parts in the developing processes. Leveraging these new data sources, we can in principle make great progress in perceiving socioeconomic status, evaluating and amending known theories, enlightening and creating new theories, detecting abnormal events, predicting future trends, and so on. Such opportunity has led to the emergence of a new branch named computational socioeconomics that studies various phenomena in socioeconomic development through analyzing large-scale real-world data by employing quantitative methods, with a particular attention to economic development problems related to social processes and social problems related to economic development [7]. The computational socioeconomics can be considered as a new branch of socioeconomics resulted from the transformation of methodology, or as a new branch of computational social science by emphasizing on socioeconomic problems [2, 8]. The main research issues in computational socioeconomics can be roughly laid out on three layers: (i) global socioeconomic development, with a particular attention to inequality [9] and fitness [10] in economic development, as well as the global evolution of culture [11]; (ii) regional economic development, with a particular attention to characterizing the status of regional economic development [12], revealing the hidden complexity of regional economic structure [13], and exploring effective ways toward better economy [14]; (iii) individual socioeconomic status, mainly concentrating on how to quantify individual employment status, financial wellness, living standard and mental condition in an efficient and unobtrusive way [15, 16].

Built on a growing number of papers mainly published in the last decade, this perspective classifies the novel data-driven methods into three representative categories: (i) natural data analyses, (ii) large-scale online experiments, and (iii) integration of big data and surveys. Each of the following three sections will introduce
one method, providing the fundamental idea, one typical research example and a few related works. This perspective ends up with in-depth discussion on the limitations and challenges of the above-mentioned emerging methods, as well as related ethical issues.

2. Natural data

Natural data refers to the data collected in a way that the targets are not aware of the fact that they are recorded and analyzed. Mobile phone data and social media data are typical examples. Natural data can best reflect the real behaviors and thus is considered to be even more honest than the so-called honest signals [17] collected by long-worn sensors, because volunteers’ behaviors will change to be more social desirable (sometimes inconscintly) if they know they are involved in some experiments.

A typical research example is how to quantify the religious segregation via analyzing social media data [18]. Religion locates in the core of culture, exhibiting significant positive values like to facilitate human cooperation [19] and to improve life satisfaction [20]. At the same time, people with different faiths tend to form relatively isolated communities, which may result in misunderstanding, prejudice, animosity, extremism or even violence [21]. Therefore, to quantify the extent of religious segregation and to find out effective methods that can promote cross-religion communications are valuable. Questionnaire is not suitable in this scenario because some questions could cause offense. Instead, Hu et al [18] collected publicly available data from weibo.com (one of the largest online social platforms in China, similar to Twitter), where users with religious beliefs (i.e. believers) are identified by natural language processing techniques (see details in [18]). The subgraph of the weibo follower–followee network, induced by these believers, is analyzed. Four major conclusions are drawn from such natural data. Firstly, users with different beliefs are largely segregated, and the religion network in weibo.com shows even higher level of segregation than well-known social networks with remarkable segregation like sexual partnerships mixed by races [22] and the Twitter follower–followee network of politicians in democratic party and republican party [23]. Secondly, cross-religion links play the most critical role in maintaining network connectivity, whose removal will lead to much faster breakdown of the network than links selected by other well-known centralities (see reference [24] for these centralities), such as betweenness [25] and bridgeness [26]. Thirdly, according to the comparison with null model [27], Buddhism contributes the most to cross-religion communications. Lastly, about half cross-religion links point to charitable nodes, suggesting that charity may be a common interest that can stride across the ideological barriers between religions.

The above work for the first time quantifies the religious segregation based on natural data. Recently, a growing number of works on socioeconomic issues have utilized natural data. For example, tweets of Christians and atheists are analyzed to show the value of religion in bringing happiness [28], searching queries in Google [29] and activity times in Twitter [30] are used to estimate unemployment, records in office automation systems are utilized to predict future promotion and resignation of employees [31], and online resumes of anonymous job seekers are used to quantify the differences in expected salaries of groups with different gender and stature, which can be considered as clues about potential discrimination in employment [32].

3. Online experiments

Although natural data is powerful in revealing underlying reality, we should be aware of that behavioral experiments are often nonreplicable. Recently, the Open Science Collaboration repeated 100 psychological experiments published in American Economic Review and Quarterly Journal of Economics in 2011–2014, as well as 21 experimental studies in social sciences published in Nature and Science in 2010–2015. In the former case, they reported a significant effect in the same direction as the original study for 11 replications (61%) with the replicated effect size being on average 66% of the original [35], while in the latter case these two numbers are 13 (62%) and 50% [36].

Using online platforms to design and implement large-scale behavioral experiments may overcome the above-mentioned shortcomes, such as high costs, insufficient samples and low representativeness. Bond et al
[37] designed an online political mobilization experiment during the 2010 US congressional elections, which attracted 61 million Facebook users. In the experiment, all users of at least 18 years old in the US who accessed the Facebook website on 2 November 2010 (i.e. the day of the US congressional elections) were randomly assigned to a social message group, an informational message group or a control group. Each user in the social message group was shown a statement at the top of the News Feed, which encouraged the user to vote, provided a link to find local polling places, showed a clickable button reading I Voted, showed a counter indicating how many other Facebook users had previously reported voting, and displayed up to six randomly selected profile pictures of the user’s Facebook friends who had already clicked the I Voted button. The informational message group was shown the message but they were not shown any faces of friends. The control group did not receive any message at the top of their News Feed. The experimental results showed that users in the social message group were 2.08% more likely to click on the I Voted button than those in the informational message group, that is, the socialized environment could promote political self-expression. In addition, users receiving relevant information in their News Feed are 0.39% more likely to vote than those in the control group. Recalling the 2000 US presidential election where Bush beat Gore in Florida by only 537 votes (<0.01% of total votes in Florida), 0.39% is more than enough to make significant difference in election. In a word, with the help of Facebook, Bond et al successfully implemented an online experiment with an inconceivably large number of participants, and showed clear evidence that online political mobilization can essentially contribute to real-world behaviors.

In addition to co-designing experiments and co-developing necessary softwares (like [37]), researchers could collaborate with online platforms in various ways. Salganik et al [38] used Internet to recruit volunteers and to launch the experimental software. They attracted 14 341 volunteers to participate in the music evaluation experiment, which suggested the existence of a strong Matthew effect (rich gets richer). Pichard et al [39] proposed a mechanism to allocate rewards to all contributors for a task, and then utilized online social networks to propagate their idea and generate the incentive network. They quickly built a team with about 4400 members and eventually won the DARPA network challenge 2009 that asks for the coordinates of ten red weather balloons placed at different locations in the continental United States. Van de Rijt et al [40] directly used existing functions of online platforms (e.g. crowdfunding website Kickstarter, online product evaluation website Epinions, etc) to implement their experiments (e.g. to randomly choose 100 projects and donate a small percentage of funding goal to them, to evaluate randomly sampled products as very helpful, etc), revealing a sharp success-breeds-success phenomenon. Pennycook et al [41] sent private messages to 5379 selected users from Twitter and recruited thousands of volunteers in Amazon Mechanical Turk to investigate why people share misinformation, and the experimental results suggested that sharing does not necessarily indicate belief but subtly shifting attention to accuracy will increase the quality of news that people subsequently share.

### 4. Survey expanding

The increasing data availability provides us unprecedented opportunity to collect and analyze population-scale data, such as social media data from Facebook, Twitter, Wechat and Weibo, and call detail records and mobility trajectories from smart phones. Although the cost to obtain such data is much lower than that of population-scale censuses or surveys, we cannot use these data to directly answer important questions concerning household incomes, mental and physical health, social segregation and social mobility, employment status, and so on. Therefore, we call these data as easy-to-access indirect data. Let us consider a certain study that requires population-scale data on household incomes. Obviously, it is very hard to know the household income of every family because a poor country cannot support a population-scale economic census, and even if the census is completed, the microscopic data is usually confidential and not open for public or research institutions. Under the circumstances, we can obtain household incomes of a tiny number of families via routine surveys. This much smaller data can be used as training data, based on which we can apply machine learning techniques to predict household income of a family based on easy-to-access indirect data, such as the mobile phone data of the family members [42]. Such a small-size data set directly pointing to the research target is called hard-to-get direct data. Combining the population-scale easy-to-access indirect data, a small sample of hard-to-get direct data, and a properly selected or well-designed algorithm to infer the missing direct data of individuals other than the sample is a novel and powerful method. Although the inferred data is not perfect, it can be obtained at a very low cost and its accuracy is usually sufficient for studying and policymaking.

Using the above method, Blumenstock et al [9] produced the wealth map of Rwanda at a very high resolution, so that poor areas and individuals can be easily identified. Collaborated with telecommunications operators, Blumenstock et al obtained 1.5 million mobile phone users’ call detail records including billions of calls and short messages. They further recruited 856 volunteers from those mobile phone users and asked volunteers to fill in detailed questionnaires about their socioeconomic statuses. They trained a machine learning algorithm to predict the 856 volunteers’ wealth and other socioeconomic indicators, and then generated...
the out-of-sample prediction for the 1.5 million users. They found a strong correlation between the government ground truth data and the predicted wealth data after aggregating them to the district level. At a very low cost (only about 0.06% of mobile users are sampled to fill in questionnaires), their method could map the distribution of wealth and other socioeconomic indicators for the full national population.

In addition to socioeconomic status, this method can also be applied in analyzing mental health conditions at the individual level. Although mental problems of specific individuals do not belong to the scope of social sciences, the temporal trends of anxiety level, depression level and suicide rate of a society, and the relationships between the overall prevalence of mental diseases and political system, economic development and educational standards are highly concerned issues in social sciences. Taking depression as example, De Choudhury et al collected 476 depression patients’ health records and their one-year Twitter data before onsets [43], as well as 165 baby mothers’ health records and their Facebook data [44]. De Choudhury et al [43, 44] respectively built machine learning models to predict depression and postpartum depression, and the accuracy of the former can approach to 70%. In a word, they used a few hundreds of health records (hard-to-get direct data) and corresponding Twitter and Facebook data (easy-to-access indirect data) to train their machine learning models, and in principle their models can be applied to the entire Twitter and Facebook data to estimate the overall level of depression of a country, a region, or a certain group of people with predefined demographics. Analogously, Weibo text [45] and Instagram photos [46] can also be utilized to identify depression patients at the early stage.

5. Discussion

Thanks to the big data and artificial intelligence techniques, papers applying the three highlighted methods appear more and more frequently in the last decade, which are leading a profound and irreversible reform of routine methodologies of social sciences [5, 7]. At the same time, there are still some apparent limitations of these promising methods, which are worth special attentions in the future studies.

In brief, a theory in natural sciences has three progressive abilities: (i) it could explain some observed phenomena, (ii) it could predict what will happen with given conditions, and (iii) it can be applied to control real systems. This explain-predict-control chain constitutes a common validation process for a newly-proposed theory. In social sciences, we cannot find or construct an idea system with every component and ingredient being explicitly and quantitatively known, and the functions and effects of a theory cannot be completely distinguished from others. Therefore, we cannot expect to understand the role of a social theory like the role of the gravity law in artificial satellites or the role of the general relativity in global position systems. Accordingly, we do not have the illusion that a social theory can be well utilized to control a real system, yet we expect at least it can be applied to intervene a system. However, in order to go from predicting to intervening, we have to know relevant causal relationships, not just correlations [33]. For example, although we know the temporal orderliness of campus lifestyle (e.g. the time to take shower) is significantly correlated with academic performance [47], and this correlation (together with some other features) can be utilized to predict academic performance with an acceptable accuracy [48], we cannot improve a student’s academic performance by simply rescheduling her/his taking shower time. It is because a regular schedule of showers is not a causation of a high GPA score, while there may be some unobservable causations (e.g. conscientiousness [49]) that lead to both high/low orderliness and high/low GPA scores. Online experiments are able to reveal clear causality and sometimes experiments themselves do intervene the real world as Bond et al did [37]. In comparison, statistical techniques that evaluate the intervention effect (e.g. propensity score matching [50] and difference in differences [51]) and discover the causality (e.g. asymmetric Shapley values [52] and causal Shapley values [53]) could provide valuable clues, but those inferred results are less reliable than experimental results. So generally speaking, we are still far from the ability of intervention. Notice that, more and more data sets with time stamps are available, which are not exactly the same to but similar to longitudinal data, therefore researchers should explore how to apply, extend and develop known tools and methods for longitudinal data [54] to deal with socioeconomic problems.

The last section introduces how big data and traditional questionnaires can be combined to get added value. In the combination, however, questionnaires are not changed. That is to say, the design of questionnaires does not benefit from big data. What we expect in the next step is that big data can help in designing better questionnaires. On one hand, routine method to design a questionnaire may miss important items, which are probably found out in advance by analyzing correlations between candidate factors and targets with the help of big data and data mining techniques. On the other hand, the gap between correlation and causality can be bridged by a properly-designed questionnaire. In this way, the novel data-driven methods and traditional survey methods are deep integrated. A very few works have already seriously investigated the feasibility and possible pathways [55]. In addition, methods to quantify variable importance that are currently used to make machine learning interpretable can also be used to understand what variables are important to a certain outcome, and this
understanding may lead to a better design of questionnaires. For example, Shapley values, coming from coalition game theory [56], has been shown to be a very good measure of global variable importance [57, 58], and thus may be applied in designing questionnaires using knowledge gained from big data.

Lastly, the methods introduced in this perspective may give rise to a number of ethical issues, which must be cautiously treated. Firstly, the private information needs to be strictly protected. When natural data is used in a study, the involved individuals are not aware of the fact that they are under investigation. If the data comes from unpublishied systems, it is clear that the privacy should be protected. Even if the data is collected from public websites, the private information has to be protected because individuals probably do not want other people knowing the analytic results related to them, or do not want to see the results themselves. For example, a user is willing to share his information to Facebook friends, which does not imply that he agrees to be known as a predicted depression patient [43] or a predicted gay [59] based on his shared data. Furthermore, even if a data set is anonymous, it is possibly to be de-anonymized [60–62]. Therefore, we should be very careful in reporting analytic results and sharing scientific data to avoid disclosure of private information. Secondly, in some online experiments, to capture the real responses of subjects, they are kept to be unaware of the experiments. Therefore, researchers have to carefully evaluate in advance whether the materials shown to subjects (e.g. rumors, violence information, extreme and fanatical views, etc) and the feedback from researchers (in particular the negative feedback against some selected subjects) will result in long-standing negative impacts on subjects’ emotion and mentality. Thirdly, both big data and survey data often contain representation bias, that is, a data set is representative of its source but may not be representative of other ensembles [63]. For example, a random sample of users from a social media website probably over-represents young people while under-represents aged people in a population. Therefore, we should be cautious when applying the trained models or extending the results to the population different from the original source. An impressive lesson is that a well-performed classifier trained on ImageNet is not good at classifying images from India because only about 2% of images in ImageNet are from India [64]. Although individual awareness of ethical issues is important and helpful, institutions and funders (may through the Institutional Review Boards) ought to play a vital role in developing informed regulatory framework and ethical guidance for researchers, and in evaluating potential privacy and ethical concerns before studies were started [8].

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grant No. 11975071.

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

Tao Zhou https://orcid.org/0000-0003-0561-2316

References

[1] Wang F-Y, Carley K M, Zeng D and Mao W 2007 Social computing: from social informatics to social intelligence IEEE Intell. Syst. 22 79–83
[2] Lazer D et al 2009 Social science: computational social science Science 323 721–3
[3] Shah D V, Cappella J N and Neuman W R 2015 Big data, digital media, and computational social science Amm. Am. Acad. Pol. Soc. Sci. 659 6–13
[4] Evans J 2020 Social computing unhinged J. Soc. Comput. 1 1–13
[5] Buyalskaya A, Gallo M and Camerer C F 2021 The golden age of social science Proc. Natl Acad. Sci. USA 118 e2002923118
[6] Fisher R J 1993 Social desirability bias and the validity of indirect questioning J. Consum. Res. 20 303–15
[7] Gao J, Zhang Y-C and Zhou T 2019 Computational socioeconomics Phys. Rep. 817 1–104
[8] Lazer D M J et al 2020 Computational social science: obstacles and opportunities Science 369 1060–2
[9] Blumenstock J, Cadamuro G and On R 2015 Predicting poverty and wealth from mobile phone metadata Science 350 1073–6
[10] Tacchella A, Mazzilli D and Pietronero L 2018 A dynamical systems approach to gross domestic product forecasting Nat. Phys. 14 861–5
[11] Schich M, Song C, Ahn Y-Y, Mirsky A, Martino M, Barabasi A-L and Helbing D 2014 A network framework of cultural history Science 345 558–62
[12] Chen X and Nordhaus W D 2011 Using luminosity data as a proxy for economic statistics Proc. Natl Acad. Sci. 108 8589–94
[13] Gao J and Zhou T 2018 Quantifying China’s regional economic complexity Physica A 492 1591–603
[14] Gao J, Jun B, Pentland A S, Zhou T and Hidalgo C A 2021 Spillovers across industries and regions in China's regional economic diversification Reg. Stud. 55 1311–26

[15] Luo S, Morone F, Sarrate C, Travizano M and Makse H A 2017 Inferring personal economic status from social network location Nat. Commun. 8 15227

[16] Di Clemente R, Luengo-Oroz M, Travizano M, Xu S, Viattla B and González M C 2018 Sequences of purchases in credit card data reveal lifestyles in urban populations Nat. Commun. 9 3330

[17] Pentland A 2019 Honest Signals: How They Shape Our World (Cambridge, MA: MIT Press)

[18] Hu J, Zhang Q-M and Zhou T 2019 Segregation in religion networks EPJ Data Sci. 8 6

[19] Purzycki B G, Apicella C, Atkinson Q D, Cohen E, McNamara R A, Willard A K, Xygalatas D, Norenzayan A and Henrich J 2016 Moralistic gods, supernatural punishment and the expansion of human sociality Nature 530 327–30

[20] Lim C and Putnam R D 2010 Religion, social networks, and life satisfaction Ann. Sociol. Rev. 75 914–33

[21] Atran S and Ginges J 2012 Religious and sacred imperatives in human conflict Sci. Adv. 8 284–7

[22] Conover M, Ratkiewicz J, Francisco M, Goncalves B, Menczer F and Flammini A 2011 Political polarization on twitter Proc. Fifth Int. AAAI Conf. Weblogs and Social Media pp 89–96

[23] Lü L and Zhou T 2011 Link prediction in complex networks: a survey Physica A 390 1150–70

[24] Zhou T, Liu J-G and Wang B-H 2006 Notes on the algorithm for calculating betweenness Chin. Phys. Lett. 23 2327–9

[25] Cheng X-Q, Ren F-X, Shen H-W, Zhang Z-K and Zhou T 2010 Bridgeness: a local index on edge significance in maintaining social cohesion Phys. Rev. E 82 056117

[26] Maslov S and Sneppen K 2002 Specificity and stability in topology of protein networks Proc. Natl Acad. Sci. USA 99 15449–53

[27] Askitas N and Zimmermann K F 2009 Google econometrics and unemployment forecasting Soc. Sci. 2 28–42

[28] Ritter R S, Preston J L and Hernandez I 2014 Happy tweets J. Phys. Complex. 2 15012

[29] Conover M, Ratkiewicz J, Francisco M, Goncalves B, Menczer F and Flammini A 2011 Political polarization on twitter Proc. Fifth Int. AAAI Conf. Weblogs and Social Media pp 89–96

[30] Llorente A, Garcia-Herranz M, Cebrian M and Moro E 2015 Social media fingerprints of unemployment Physica A 431 854–6

[31] Yau H, Lian D, Cao Y, Wu Y and Zhou T 2019 Predicting academic performance for college students: a campus behavior perspective J. R. Soc. Interface 15 20180210

[32] Yang X, Gao J, Liu J-H and Zhou T 2018 Height conditions salary expectations: evidence from large-scale data in China Physica A 501 86–97

[33] Pearl J 2009 Causality: Models, Reasoning, and Inference (Cambridge: Cambridge University Press)

[34] Open Science Collaboration 2015 Estimating the reproducibility of psychological science Science 349 aac4716

[35] Camerer C et al 2018 Evaluating replicability of laboratory experiments in economics Science 351 1433–6

[36] Chan C, Wang Y, Lam H M and Zhou T 2017 Blockchain: a local index on edge significance in maintaining social cohesion Phys. Rev. E 96 042308

[37] Diggle P, Heagerty P, Liang K Y and Zeger S 2013 Analysis of Longitudinal Data (Cambridge: Cambridge University Press)

[38] Blumenstock J E 2016 Evaluating replicability of laboratory experiments in economics Science 351 1433–6

[39] Rosenbaum P R and Rubin D B 1983 The central role of the propensity score in observational studies for causal effects Biometrika 70 47–55

[40] Abadie A 2005 Semiparametric difference-in-differences estimators Rev. Econ. Stud. 72 1–19

[41] Frye C, Rowat C and Feige H 2020 Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability Advances in Neural Information Processing Systems vol 33 pp 1229–39

[42] Wang X, Zhang C, Ji Y, Sun L, Wu L and Bao Z 2013 A depression detection model based on sentiment analysis in social media Algorithms 6 327–30

[43] Pennycook G, Epstein Z, Mosleh M, Arechar A A, Eckles D and Rand D G 2021 Shifting attention to accuracy can reduce misinformation online Nature 592 390–5

[44] Pearl J 2009 Causality: Models, Reasoning, and Inference (Cambridge: Cambridge University Press)

[45] Lundberg S M 2020 From local explanations to global understanding with explainable AI for trees Nat. Mach. Intell. 2 56–67
[59] Kosinski M, Stillwell D and Graepel T 2013 Private traits and attributes are predictable from digital records of human behavior Proc. Natl Acad. Sci. 110 5802–5
[60] Wondracek G, Holz T, Kirda E and Kruegel C 2010 A practical attack to de-anonymize social network users Proc. 2010 IEEE Symp. Security and Privacy pp 223–38
[61] De Montjoye Y-A, Hidalgo C A, Verleysen M and Blondel V D 2013 Unique in the crowd: the privacy bounds of human mobility Sci. Rep. 3 1376
[62] De Montjoye Y-A, Radaelli L, Singh V K and Pentland A S 2015 Unique in the shopping mall: on the reidentifiability of credit card metadata Science 347 536–9
[63] Suresh H and Guttag J A 2019 Framework for understanding sources of harm throughout the machine learning life cycle (arXiv:1901.10002)
[64] Shankar S, Halpern Y, Breck E, Atwood J, Wilson J and Sculley D 2017 No classification without representation: assessing geodiversity issues in open data sets for the developing world (arXiv:1711.08536)