SOCG 290: Personal Report

Name: AMIT BORASE
Pid: A53095391

Traditionally sociology has been the field of qualitative research. This has been partly due to the lack of substantial and veracious datasets to model the research on. Data collection techniques in sociology has been limited to representative sampling through participant observation, content analysis, interviewing, and documentary analysis. Such studies have often been very slow and give delayed results. Advent of big data poses a serious alternative to such traditional ways of research in the field of sociology.

Paradigm Shift with Big Data

The rise of the Internet, social media, and digitized services has produced a colossal amount of data in recent years. Such 'big' data offers the possibility of access to social aspects such as interactions, behaviours and opinions at a scale beyond the dreams of sociologists who had to settle for traditional representative sampling. Big data is often described as being characterised by the '4 Vs': volume (the large scale of the data); variety (the different forms of data sets); veracity (quality of the data) and velocity (real-timeness, the constant generation of these data). Social scientists are now using huge data sets, many produced through online interactions and media, that shed light on basic social processes. Big data data sets, from sources like Twitter, Facebook, or mobile phones, give social scientists ways to tap into interactions and cultural output at a scale that has never been seen before in social science. The way we analyze data in sociology and organizational theory are bound to change due to this explosion of new data.

Social Big Data Analysis

The advent of big data has been so impactful for field of sociology that it has extended into new branch called 'Social Big data analysis', additionally it has shifted research in sociology from qualitative methods to quantitative ones. Computer scientists have produced powerful new tools for automated analyses of such big data, but they lack the theoretical direction necessary to extract value from such analysis. Meanwhile, cultural sociologists have produced sophisticated theories of the social origins of meaning, but they lack methodological capacity to explore them. Synthesis of these two fields that adjoins conventional qualitative methods and new techniques for automated analysis of big data poses as great opportunity for sociology as a whole.

Social data analysis mainly involves two-step processing: first large quantities of data is extracted in meaningful, analysable formats from online platforms using sophisticated computer science machinery. Such data is then quantitatively analysed with careful weighing-in of factors such as influence, reach, and relevancy, an understanding of the context of the data being analyzed. In short, social data analytics involves the analysis of social media in order to understand and surface insights which are embedded within the data. Social data analysis has applications in wide areas such as marketing and advertising, law and enforcement, health care services, politics, demographics, education, recruiting, banking, stock markets etc. Below are some of the applications of Big data that have immediate value in the field of sociology.

Sociological Stratification Using Big Data

Pierre Bourdieu's classical theories on social stratification have been widely studied by the sociologists. Bourdieu proposes that the aesthetic choices of a person create *class fractions* (class-based social groups) and actively distance one social class of the society from the other. These choices are often the result of the socioeconomic background of the individual in society. Advent of Big Data provides us a renewed opportunity to revisit this theories, but now equipped with more accurate and larger dataset. Big data is a compelling tool in studying these theories because platforms such as social media, online shopping, have permeated into each and every classes of the society irrespective of economic or cultural biases, hence big data generated from these platforms embodies characteristics unique to each social class. Additionally datasets thus obtained matters for social science in two different ways. First, they have varied implications on inequalities, democratic participation, the distribution of wealth. Second, they offer new methods to be exploited to gain insight into a wide range of traditional and evolving social phenomena, such as consumer behaviours, political endorsements and etc.

We set out to explore such sociological implications of big data analysis in our first project, by undertaking a Big Data approach in trying to 'make-sense' of Amazon raters using Bourdieu's theoretical concept of class distinction and taste. Online retailer such as amazon and its recommendation systems serve as rich repository of big data that can be analysed and investigated to stratify customers/raters based on their socio-economic backgrounds both in terms of cultural taste and economic freedom. We chose headphones as the subject of our case study because products offered in this segment vary greatly in their technical and fashion centric capabilities and hence target various consumer stratas of the market. We began with the hypothesis that the buying behaviours of the headphone consumers will greatly follow the cultural and economic backgrounds of the individual. We targeted headphone products from mainly 3 price brackets – lower (10-45 dollars), middle (50-140 dollars) and upper (175 dollars +) . This was followed by a deliberate scraping and sampling of review data of 15 headphones, wherein inconclusive (shorter) and biased (ex. manufacturing defects, logistics etc) reviews were filtered out. We proceeded by tokenizing review data followed by decomposing it into collection of bigrams on which LDA topic modeling was performed to generate 40 topics. These topics were then manually analyzed and further filtered to obtain classification of reviews into 2 groups – Audio focused and Fashion focused. This is done using sentiment analysis of review data. Using sociological theory, this groups were further subdivided into 2 categories for each class, Fashion focused into yuppies and disenfranchised and audio focussed into pragmatists and purists. Pragmatists represent consumer strata high in economic and cultural capital, while the purists represent the ones with higher cultural but lower economic capital. Similarly the yuppies represent the class with higher economic but lower cultural capital and disenfranchised represent those which lack both in cultural and economic capital. The results thus obtained were keeping in-line with our hypothesis. The word-clouds formed for each class closely resembled the economic and cultural characteristics governing the same.

Big Data to Analyse 'Social Arenas' & Their Evolution

Modern society is governed by strong social structure, that refers to the network of interrelated roles and statuses that guide human interaction. People interact with each other in one way or the other, almost every day. These interactions can be of various types such as cooperation, competition, exchange, conflict and accommodation. Digitization of processes and resources means, many of these interactions happen over automated platforms. Data generated from such platforms holds key in studying patterns in these interactions. Such interactions can often be encoded in the form of sequences , which can then be churned through statistical machinery of sequence analysis and optimal matching. Social sequence analysis forms a group of such statistical techniques, that facilitates study of interpersonal contact dynamics, development of social hierarchies and macrosocial temporal patterns.

Sequence analysis has powerful implications when it comes to analysing sociological evolution. We explored this area in our second projects to analyse patterns in the career trajectories of american sociologists, using techniques such as sequence analysis. Objective of this work was to find out patterns in the publications and organizational associations of the academicians in the field of american sociology. This interest was result of rise of quantitative analysis based research in the field of american sociology. We chose web of science repository as our primary source of data for analysis. We scraped sociological publication data based on multiple filters such as topic, field, journal, language and region. Such data was arranged in dictionary format. Further the universities and journals were manually classified into different tier based on their reputations and impact factors. This information was encoded back into the dictionary data, which was then run against the K-medoid clustering algorithm using the optimal matching distance functions, to obtain the clusters.

Patterns thus generated allowed us to find out how academic careers evolve. Results suggested that the number of publications by an individual tends to be a lower number, and majority of the publications are made to the lower tier journals while only few publications fall into top tier journals. Moreover, it is the initial set of publications that greatly defines the overall career. Additionally we found out that the scholars tend to change universities very rarely and tend to stay in university of their first publication. All of these patterns are very obvious and not significantly surprising. This could be due to some inaccuracies and limitations in the dataset, arising from the difficulty in scraping data from WOS. Major learning from this project for me is that the big data repositories such as WOS have built various restrictive frameworks to guard themselves against the unwanted access to their data. As a researcher, it is very important that you devise novel techniques and hacks to get around these.

Sociological Interactions and Big Data

Social networks have always been the central theme of the sociological evolution of the humanity. Networks are just a systems of social relationships that can exist in any social context such as the family, school, workplace, village, sports-team and so on. So it is quite inaccurate to suggest that the social networks are phenomenon of the 21st century. Internet platforms such as Facebook and Twitter are merely the channels of these relationships, and have greatly simplified the environment for the networked interactions to occur. Major shortcomings to study 20th century social networks arise from the partial lack of substantial ('big') and accurate data governing interactions in these networks. This is where Big Data from social network platforms has lot to offer, because now sociologists have easy access to substantial quantities of data that signify the networked interactions between various classes of society. Additionally modern day computational capabilities further allow Sociologists to take entirely quantitative approach to the research by analysing such data in greater depth to identify sociological patterns. While embracing the quantitative approach with Big Data, we have to be careful by not neglecting qualitative aspects entirely. "Social" networks often revolve around the world of meanings, feelings, relationships, attractions, dependencies, which have traditionally been at the heart of qualitative research.

On the similar lines, our third project involved studying the nature of interactions among entities of social networks. We started with a goal of identifying and visualizing the collaboration patterns among the scholars in the field of american sociology. Again we chose WOS as the repository for the publication data of sociology scholars, and directly extracted data from the website. Data so obtained lacked the institutional affiliations and tiers attributes of the authors. Hence we had to code a solution that scraped authors and their institutional affiliations along with other metadata informations separately from WOS and encoded the same back into the data downloaded directly from WOS. We then fed this modified data into the Sci2 tool to build a co-author network from the given data. Followed by this we used popular visualization tool called Gephi for efficient visualization of co-author network, we had to do additional tuning to the network parameters of gephi to be able to obtain meaningful network graph. We were able to make some straightforward observation such as cross-tier collaborations are very rare, in other words authors tends to collaborate with those that publish in the same tiered publications. The rare cases of cross-tier collaborations may arise from situations such as collaboration between former student and advisors. Overall we concluded that the class distinctions are central to the system of collaboration networks.

Challenges

Below are some of the challenges that needs to be tackled to be able to transform big data analysis as potent research tool for sociology and science as a whole.

- *Qualitative Value of Data:* As a practitioners of social data analysis, we have to be careful when it comes to framing studies. Analysts often leave out or fail to identify the qualitative value of the data set, and perform only quantitative analysis on the data. If big data studies are going to take over the field they need to address pressing theoretical problems, rather than provide models that just tell the evident story. This can be achieved by facilitating dialogue and collaboration between the social sciences and computer science around big data and their use for the advancement of knowledge, policy, and more generally society.
- *Privacy Invasion:* No discussion on Big Data could be complete without mentioning the increasing concerns about privacy. The data we choose to extract may not always be available openly, or in some cases available but not lawfully. Such datasets are often exclusively owned by or are private property of individuals or organizations. In such case, we have to be aware of the lawful implications of harvesting such data and respect the same.
- *Causation from Correlation:* We have to be careful about *Inferring Causation from Correlation.* Big data is powerful tool for predicting sociological behaviour pattern, but should we really get lured by it so much so that we use it to base some of the basic decisions pertaining to human life? For example, how right is it for parole boards to make judgments on release of incarcerated convicts solely based on the analysis performed on social big data, which may not even be relevant to the concerned individual.
- Technological Needs: Data generated from various social networks platforms is of phenomenal scale (facebook currently hosts 1 Billion+ users), this is only going to increase in the future. Analysing such data is computationally demanding task. We have to come up with technological advances to be able to sustain such computational needs.

Conclusion

Big data offers tremendous potential in the field of sociological research. It's easy accessibility, large size, diverse nature and current-ness make it very attractive resource for social analysis, to unearth patterns in human behavior and interactions, especially in the areas of sociology related to social stratification analysis, sociological evolutional patterns and social network analysis. Big Data at its heart encodes human qualities such as emotions, feelings, perceptions and thoughts. Hence it is very important that the traditional ways of qualitative analysis are not forgotten completely or missed while performing quantitative analysis of such Big Data.

Acknowledgements

I entered this course with limited understanding of the Big Data revolution and its overall implications on the field of sociology. I can confidently say that the project work done and discussions undertaken in this course furthered my knowledge of Big Data phenomenon and its sociological implications. I thoroughly enjoyed the hands-on nature of the course and thank Prof. Guerra, Apoorva and my team members who helped me in various ways during the entire course.

References

- SOCG 290: Course material.
- www.wikipedia.org
- https://databigandsmall.com/tag/sociology/