STRIKE 1st Stakeholder Workshop
4th of September 2024, LSHTP

Introduction

The STRIKE project is an ESRC investment as part of the Transforming data collections infrastructure for social science call. The project sets out to scope the possibility to produce a nationally representative dataset of social connections. The workshop was a first engagement to scope what kinds of data stakeholders might want or need and what this data could be useful for. The workshop consisted of three presentations, Gilbert on an overview of STRIKE, Edmunds on Challenges to collecting social contact data for epidemiology and Collins on innovations in data collection methods. Most of the workshop consisted of discussion, starting off with blue sky session on what connections data might be used for, later going more into problems of collection, ethics, democratic accountability etc.

Participants showed high levels of interest in the data. The stakeholders were a mix of social scientists, epidemiologists, methodologists, social networks researchers across academia and the public sector.

Interest in the Data

There was a lot of interest in all kinds of connection, connectivity, relationship and network data. Application areas touched on were infectious disease epidemiology, children online safety, social capital, better understanding behaviour change, networks and commercialisation, understanding life-course network changes, contacts and networks within institutions (care homes, prisons).

Infectious disease epidemiology

For the epidemiological application the discussion focused on two aspects. The first extension to existing contact data for better understanding infectious diseases discussed was to manage to collect socio-demographic data alongside the connections. Whilst the who-meets-who according to age is fairly well researched and documented, who-meets-who across ethnicities, social class, socioeconomic group or even gender (is that right?) is largely unknown. The second extension is to integrate the epidemiology with the economy. For example, whilst we could predict the infection outcomes of people working from home, these findings were not integrated with modelling the economic consequences, e.g. for pubs in high density office areas.

Online and offline networks, safety and social capital

One application area highlighted was the importance to better understand children's online social networks. Online data was also discussed in terms of social capital and social network data. One problem highlighted was the unavailability of online network data which is held by private corporations that have no interest in releasing

Networks, Behaviour Change and Commercial Use

Networks, connections and connectivity was identified as one driver of behaviour change. People are more likely to adopt something when they see it in their neighbours and peers. This holds for positive aspects, eg. adoption of solar panels (highly clustered across cities identifying imitation across neighbourhoods) as well as negative ones, e.g. smoking or vaping. There was a clear

danger of exploitation of network and connectivity data for commercial use, e.g. recommender systems.

Institutional Settings

Institutions were seen as potential settings to extract networks and connections from. Data can be extracted from work flows and rotas. Understanding connectivity in settings like prisons and care homes has potential to better understand outcomes of public policy and value for money.

Cross Sectional or Longitudinal Data

Questions of what kind of data collection would be needed, cross sectional, repeat cross sectional or longitudinal. One use for longitudinal data collection identified was to better understand how networks change over the life course, in particular across life course events - school, university, work, marriage, children etc. There was distinct interest in this kind of research across participants. The interest was first of all to understand what changes to contacts and networks occur across the lifecourse and also to better understand the drivers of those changes.

Collecting Social Connection Data

Infection data

There are many different ways to measure contacts, but all of them have problems.
Diary methods rely on memory and most likely leave out several contacts that have fallen out of a person's recall. Proximity measures will record spatial snapshots but most likely leave out several connections between places. Location based estimates will most likely overestimate contacts, producing complete networks between colocated individuals. Direct observation data is potentially the best collection method but is intrusive (CCTV), potentially unethical (surveillance without consent) and expensive (equipment). Data can be triangulated (eg google mobility data with infection risk data). Existing data collections lack covariate data and are often focused on respiratory infections, with other transmissions neglected.

Methods

Traditional methods for collecting network and contact data like surveys and observation studies are possibilities for the STRIKE infrastructure. Surveys are, however, a relatively high burden on the participants and suffer from recall failure etc. There are surveys which can be leant on in developing the connections survey eg. Time Use surveys, National Diet and Nutrition survey, National travel survey.

Apps are a new way of potentially collecting connections data. As a passive collection method they greatly reduce the cognitive load and allow collection of data at scale. They do, however come with their own problems of digital exclusion, measurement issues, like errors and missing data. There are also data interpretation challenges (app based methods collect enormous amounts of data which needs to be interpreted sensibly) and have their own ethical and legal challenges (surveillance).

One fruitful way forward discussed was the use of secondary data, discussed in the next section.

Integration, linkages and Generation

The first way to use existing data is to integrate contact questions into existing data collections, eg. Understanding Society, Birth Cohorts. It seemed unlikely that US would be keen to add more questions due to already being a long survey and battling non-response and attrition.

One area to explore is to see how dataset linkages can be exploited to better understand social contacts. For example, linking admin datasets and/or commercial datasets (store card data) one might be able to generate colocation and contact data.

AI was suggested to generate social contact data from exisitng data, e.g. social media data.

Ethics and Democratic Legitimacy

Throughout the day there were questions about ethics, consent and democratic legitimacy of the data. One ethical problem in social connectivity data is the identification of individuals as contacts that are not part of the study (and thus have not given consent). A second ethical problem is the intrusiveness of the methods and the level of potential surveillance exerted by them.
The surveillance aspect also lead to a discussion about the democratic legitimacy of the data collection. There are real questions about the dataset and its safety but also the collection method and its appropriation by malign actors.

Summary

A question about participation can be summed up by the suggestion that we need to build the SocialBank - the social science pendent to the BioBank, where participants voluntarily submit their network and contact data, with complete trust in its safety and usefulness.

The workshop raised more questions than it answered.
Figure 1: Heterogeneity (x-axis) and meaningfulness (y-axis) of contact/connections/network.

Discussion of the challenges of collecting contact/connections/network data has lead to questions about how much heterogeneity is needed as well as how far we can go up in therms of intensity or meaningfulness of the contacts.