Chapter 3

Data ethics

Scientists must be ethical and conscientious, always. Data bring with them much promise to improve our understanding of the world around us, and improve our lives within it. But there are risks as well. Scientists must understand the potential harms of their work, and follow norms and standards of conduct to mitigate those concerns. But network data are different. As we discuss in this chapter, network data are some of the most important but also most sensitive data. Before we dive into the data and methods in later chapters, here we discuss the ethics of data science in general and network data in specific.

3.1 Introduction to data ethics

When working with network data, we must be keenly aware of a multitude of ethical issues. Although complex ethical issues can exist for *any* data, network data poses additional challenges because individual data points are not isolated from each other. So let's walk through the common challenges as well as more network-specific issues.[1]

The very first thing we need to understand is that a dataset may contain deeply private data and the privacy may be due to the relationships contained in the data. How would you feel if all your social media activities were sold and shared? How would you feel if your personal genome or health records were accidentally published? When dealing with data about social relationships and activities, privacy issues become even more thorny because it is not only about the information and people captured in the dataset, but it extends even to the information or people who are *not in the dataset*. For instance, capturing communication activities of my friends may well include lots of information about myself. So what are the important privacy risks? What should we do to mitigate those risks?

Another important fact is that *a dataset is never an object reality*. Because it is not possible to capture *everything*, data collection always forces choices of what to

[1] If you have not done so already, we strongly encourage you to familiarize yourself with human subjects research ethics and practices, such as informed consent and institutional review boards (IRB).

https://doi.org/10.1017/9781009212601.005 Published online by Cambridge University Press

collect, what to ignore, and how. Such choices are often unintentional but their knock-on effects can introduce biases and harms. For instance, consider genetic data. Because the data collection is primarily done, particularly during the early days of genetics, by elite research universities and prestigious hospitals that are engaging in cutting-edge research, the samples tend to come from *rich areas of rich countries*, making the composition of the sample largely focused on those with European ancestry [241, 434]. What could be the implications of such biased omissions and inclusions?

Finally, ethical considerations extend beyond what is in the data itself to include how the data is processed, used, shared, and published. Data can be misused in many ways, ranging from carelessness to outright research misconduct. The misuse can start simply from carelessness—not being aware of the inclusion bias, privacy risks, and other ethical issues. Even without any bad intentions, such carelessness still can produce social harm. For instance, not fully understanding the data may lead to erroneous analysis and biased conclusions, which can then lead to problematic social policies or medical practices that have lasting impacts on society. We also hear about data breaches all the time; it is not trivial to protect the privacy of people captured in the dataset. Finally, data can also be maliciously manipulated—for example, due to the pressure towards "better" or "cleaner" results in academia—by researchers. What should we do to prevent such misuses, errors, and misconducts?

3.2 Biases in the dataset

Inclusion bias

Creating a dataset requires making choices and those choices can introduce systematic biases. At the first glance, this may not sound too bad. Why is it a problem to have fewer people from a certain ethnic group or gender? Datasets always capture only a small sample of the population anyway, right?

Here is the problem: because models, insights, and policies are derived from data, and if the data do not represent a certain population, those models, insights, and policies will not represent the population either. Here are some stories.

Did you know that when drugs are withdrawn from the market due to health risks, those health risks tend to be greater for women [209]? Although it is difficult to pinpoint the exact reasons why this is the case, an important context, and a likely culprit, is that there has been a severe sex-inclusion bias in biomedical research. Although there are numerous sex-based differences from cellular to physiological functions, testing both sexes has not been a consistent practice. If experiments are performed only on male cells, male animals, and male humans, it is only natural to expect that we cannot exactly know how a drug will work for female bodies. Although this issue was recognized and policies have been implemented requiring researchers to include both sexes in their studies, the practice is still far from perfect [449].

Another example is facial recognition. A team of researchers found that commercial facial recognition engines are terrible at detecting faces of Black people [83]. In a stirring presentation, the lead author Joy Buolamwini showed a scene where her face was not recognized by a computer *until she puts on a white mask*. This is again rooted

in the inclusion bias of the dataset. Facial recognition models are trained on large databases of human faces. Because it is much easier to find portrait photos of those who live in rich countries (Who have more cameras? Who have more websites?), these image datasets tend to include more "white" images than other ethnic groups, particularly Black people.[2] Guess what kinds of faces are well represented in such databases? Guess what kinds of faces are most accurately detected by the existing facial recognition models? What will happen if self-driving cars, drones, or other machines cannot properly recognize Black human faces as people?

Let's return to the example of genetic data. Inclusion bias in genetic data goes back to the most fundamental genomic data—the human genome. After the Human Genome Project was completed, some researchers and doctors began to discover that they couldn't match some of their patients' (especially those without European ancestry) genetic sequences to the reference genome, which hampered their diagnoses and treatments. Why did it happen? Well, you know the answer: biased sampling.

The African population is collectively much more diverse than any other subpopulation and this biased sampling of human subjects ended up creating a systematic bias in the reference human genome. A study examined 910 people and found that 300 million letters of DNA were missing in the reference genome [426]. Using one or a few individuals as genetic reference for the genome simply cannot represent the breadth of the entire human population.

We see the same pattern over and over again across domains. In addition to the cases of computer vision (facial recognition) and genomics, natural language processing (NLP) and understanding (NLU) have been criticized for inclusion bias as well [50]. Because the web is the primary source of natural language examples, the training data is heavily biased towards the dominant web platforms, languages, and populations who control the web—the same bias again.

In fact, from psychology, we already have a nice name for these sample biases: "WEIRD"—"Western, Educated, Industrialized, Rich, and Democratic." This acronym was coined to capture the most common biases in psychological research. Because a huge fraction of psychological research has been conducted in the universities of Western countries, the easiest subject to recruit for university researchers are students in those same universities, who tend to be WEIRD. In other words, a lot of our "understanding" of human psychology may be about the human psychology of the "WEIRD" population, which may or may not generalize to the rest of the world.

Any algorithms or methods trained on biased data will also be biased. Increasing the size of the dataset does not necessarily solve the issue as long as the source of the data is already biased. For instance, let's say we are training a machine learning model based on various personalized genetic and biological interaction data. Just like a facial recognition model failing on Black faces that were not in the training data, our precision health model will not know what to do with the types of data that were not in the training dataset. This will likely lead to worse performance on the underrepresented population in the dataset, which will directly lead to ineffective or even dangerous results.

[2] Even basic photographic technology has a historical bias against dark skin tones [272]!

https://doi.org/10.1017/9781009212601.005 Published online by Cambridge University Press

Data reflect systematic biases in reality

Suppose our dataset does not suffer from any inclusion bias (not realistic). Is it then free from ethical issues? Unfortunately, the answer is still no. Simply reflecting reality is still problematic because society suffers from systematic biases. If a machine learning model learns and reflects exactly what is in society, it is completely natural to see the same societal biases emerging out of the model, *even if the data does not have inclusion bias*. For instance, it was discovered that translation services such as older versions of Google Translate produced highly stereotypical results when asked to translate sentences that involve gender stereotypes. In one case, when asked to translate the ungendered Turkish phrases "O bir doktor. O bir hemşire" (He/She is a doctor. He/She is a nurse) into English, Google happily transformed this un-gendered sentence into a gendered sentence: "He is a doctor. She is a nurse."[3]

One may argue that this is not a biased model because it simply reflects reality. The problem is that a biased model can further strengthen societal biases and stereotypes. Furthermore, even a small bias can be amplified by the model; if a model is set up to return the "best" answer, even a slight tilt (e.g., 51% male vs. 49% female) can lead to the case where the model returns "male" every time!

Another issue is about biased measurements. Even without inclusion bias, the measurement itself can be biased due to systematic biases that we have in our society. A recent study demonstrated that a widely used, commercial healthcare decision-making algorithm exhibits a serious bias against Black populations [351]. This algorithm is trained using the *amount of healthcare cost* that a patient will incur as the outcome measure. This sounds reasonable because the total healthcare cost should reflect how bad a patient's health is. If they are healthy, they don't need to incur much cost; if they are seriously ill, it will naturally involve more visits, tests, drugs, and procedures, which will lead to a higher cost. The problem is that healthcare exhibits systematic biases against Black patients. Black patients have been historically marginalized, abused, and neglected by the healthcare system in the USA and that means they may incur, ironically, lower cost—by not being properly treated—than white patients given the same condition. And this is exactly what the algorithm picked up. Given a similar condition (how sick a patient is), the algorithm produced lower risk scores for Black patients (because they are likely to incur lower cost in the future), and thus recommended weaker interventions and treatments. When this type of biased algorithm is widely used (it *was* already), it will exacerbate the existing systematic bias.

Thinking about biases in network data

But you may ask, "but these examples are not about network data, aren't they?" Well, although they may not seem directly related to network data, remember that many network datasets are built on primary data such as genomic data, social data, and so on. Any bias in the base data can creep into the network data. Even worse, we may overlook the serious biases because they are not obviously visible in the network data. Once we

[3] Google's translation service has since improved its handling of gender-specific alternatives. Other machine translation services still reflect this bias.

get the "abstract" network data, we often pay little or no attention to how the network was obtained.

Working on biological networks? Any biological networks that we construct and analyze rely on, in one way or another, the human genome, or some biological samples that are quite likely biased. Are you studying protein–protein interaction networks? From whom were those interactions captured? What were the biases in the subjects who donated their blood or tissues? If they are heavily biased, can you guarantee that your methods safely generalize to other underrepresented ethnic groups? How about brain networks? Who were the subjects that your data were built on? Are you making claims about human social networks? What kinds of data are you basing your claims upon? Are the networks that you are considering heavily biased towards a specific population of European–North American subjects? Is your data coming from only the "WEIRD" population? What can you say about the rest of the world?

What should we do?

Unfortunately there is no easy solution to these issues. Although these biases can have a profound impact on subsequent analyses, it is not always easy to mitigate—or even detect—them. Furthermore, it has not been a common practice to examine, investigate, and disclose potential biases when a dataset is being collected and released. However, there is growing consensus and systematic efforts to mitigate biases in the data.

One important initiative, and a critical step, is about better documentation (see also Ch. 18). For instance, establishing a norm to always have a *datasheet* that explicitly details crucial information about the data and data collection process, including the potential biases in the data [177], will be a meaningful step. Completing a formal statement on the biases in the data, ideally from the very beginning when the data collection is planned, can help researchers to recognize biases and mitigate them, or at least let users be aware of the biases in the data.

Another important movement is the call for methods and policies for *algorithmic audits*. Algorithmic audits are the practice of systematically probing for biases in the data (and algorithms that use those data) just like auditing financial books. Good examples are the stories mentioned above, such as the facial recognition case and the case of algorithmic bias in healthcare [351]. Although an audit is not guaranteed to identify problems in the data or algorithms, it can ensure better transparency and accountability.

Finally, the development of models that can de-bias or mitigate existing biases is another active area of research [429, 16, 457].

In sum, we need to think about data bias all throughout the life cycle of datasets and their usage. It should start from careful planning that considers potential biases drive by data collection—for example, we need to ask, "what kinds of biases are we introducing by recruiting people locally from our town or university?" Further questions should be asked about measurements. Does this measurement reflect systematic biases in our society? These considerations should then be clearly documented in a datasheet that accompanies the dataset. Even if data collection is imperfect (which it will be for most cases), clearly documenting and communicating why and how can prevent downstream misuses and misinterpretations of the dataset. We should also ensure that any models trained on the dataset should provide, especially when used for important

decision making, transparency and interpretability, by using more interpretable models, by allowing independent external audits by third parties, and by publishing necessary details of the models. These steps can only be realized when we are keenly aware of the mechanisms and implications of the aforementioned ethical issues.

3.3 Privacy and surveillance

Some call the era we are living in "the age of surveillance capitalism" [512], where personal data is obsessively collected, commercialized, weaponized. Unfortunately, network data is at the heart of these practices. Let us talk about the issues of privacy and surveillance regarding network data.

What is special about network data?

Just like other data, network data poses privacy risks. But network data can pose even more risks because of the *connections* that it contain. For instance, imagine a dataset that contains the network of sexual relationships, which was collected to study the sexual behavior of people as well as the spread of sexually transmitted infections. If you are one of the "nodes" in the dataset, would you be OK if your identity was revealed? Probably not. OK, assume that your identity will *not* be directly revealed. Are you OK with the information of other people being revealed? How about the identity of those who are *connected to you* in this network? It can be pretty easy to identify you, once we know the other nodes that are connected to you.

As you can see in this example, in the case of network data, privacy is not just about *yourself*. Due to the connected nature of the network data, it is also about the data of other people, especially those who are connected with you, in the network. Even if *you* did not agree to share your "data" about your past sexual relationships, your previous partners may be willing to share, disclosing highly sensitive, revealing information about you.

Public information should be OK, right?

You may say this is an extreme example that deals with intimate personal information. How about public information? Are you OK with sharing all of your information that is "public"? If you say yes, you may want to rethink your answer after reading about a study conducted on Facebook's public "like" data. Researchers found that, if they have access to a large set of *public* data about what people *liked* on Facebook, they can reliably predict all kinds of *private* attributes of people in the data, including their sexual orientation [251].

How about some stupid, sarcastic tweets that you wrote many years ago and completely forgot about? Are you sure that you have *never* offended anyone—who were *wrong* of course—on the Internet? Even if the information is *in principle* public, it carries a different weight when aggregated into a dataset that can be systematically analyzed, searched, and scrutinized. Furthermore, once combined with other data, even

public information can have a devastating power to ruin one's life. Digital footprints, albeit public in principle, can haunt us after many years.

Limits of anonymization

Then, you may say, "OK, but it is fine as long as the data is anonymized, right?" Unfortunately, it has been demonstrated that safe anonymization is extremely challenging when social data is involved. Here is a story. The IJCNN (The International Joint Conference on Neural Networks) set up a challenge using a social network dataset, by partnering with Kaggle, a well-known platform for machine learning competitions. The challenge was to predict missing links (Ch. 10) in a social network of users of the Flickr online photo-sharing service. The data were prepared by *anonymizing* it, stripping away all user identity information, and then removing a fraction of the links between users which competitors then tried to predict.

Can you guess how the competition was won?

The winning solution *did not perform "link prediction"* per se [327]. But then how could they accurately predict links without predicting links? Instead of predicting links, they crawled the Flickr social network data *themselves* and then matched it to the competition data, effectively *deanonymizing the competition data.* They used a common property of real networks—a highly heterogeneous degree distribution. Because there are only a few nodes with large degree, they are fairly unique and identifiable. Once you match those *hub* nodes, they can act as reference points to match other nodes. In doing so, the winners could reliably match the networks even without any identifying information, and from there simply *identify links missing from the competition data* that appeared in the network that they had crawled. You don't need to predict the links if you already know them!

This "solution" was an important lesson on just how easy it is to *de-anonymize* social network data. Even without any identifying information attached, it is still possible (and fairly easy) to recover the identity of the users in the dataset if you have access to the full dataset. That means, if a network dataset with sensitive, private information is shared, *even if it is anonymized*, it may still be possible to reidentify the users in the data by cross-referencing publicly available data and then linking private information to the identified users.

Also, as mentioned before, when the data is about the social network, it is not always possible or easy for an individual to opt-out from the data release. An interesting example is the search for Saddam Hussein after the 2003 invasion of Iraq. US military intelligence collected detailed data about the social network around Hussein, to identify the most likely connections that he might have among his kins and close allies, directly contributing to his capture in December 2003 [390]. This case illustrates the power and danger of social network data and analysis. Even without any direct consent from you, the data *surrounding* you can be collected and leveraged against you.

Major features of social networks make it difficult to anonymize any social network data. First, as mentioned, the network around us is not *homogeneous*. Some people have far more social connections than others. This heterogeneity makes a social network extremely easy to "identify." Second, social networks exhibits strong *homophily*, the phenomenon that people tend to share attributes with or be similar to their social

connections. This is due to a host of mechanisms, from genetic inheritance to social contagion and population sorting. In other words, people around you may share similar genes, become similar with you because you talk with them, or you may be close to them because you are similar to each other. All of this contributes to privacy or the lack of it: if someone knows a lot about your friends, they also can infer a lot about you.

Surveillance capitalism

These privacy issues are difficult to regulate and mitigate. Companies, with strong pressure to squeeze out more profits, will collect as much data as regulations allow. And the more individual, private information the company has, the more money they can make from it, because they can more precisely target the population. Similar pressure exists for governments. The need to provide citizens with safety and security in the face of threats, likely or not, will lead to more surveillance. With better technologies and online-based communication, governments around the world have increasingly stronger power to monitor their citizens. NSA's PRISM program may be the most famous example, although similar programs exist across many countries. These circumstances and systematic pressures are often captured by the term "surveillance capitalism" [512]. The more surveillance a company does, the better they understand and predict human behaviors, which leads to more ways to make money out of them.

Open science and privacy

One area where privacy issues directly collide with another principle is in scientific research. Faced with the widespread replication crisis and other problems such as scientific misconduct, scientists and funding bodies are compelled to push for open science, sharing data and methods publicly so that other researchers can more easily reproduce and build on published studies. But then what should we do with sensitive, private data? Sharing can pose a serious privacy risk; *not* sharing violates the best practice of open science.

This again, just like all other ethical issues, does not have an easy solution. It is impossible to have strict guidelines that can be applied to every case because each case may have completely different risks and benefits when sharing the data. When the risk of deanonymization and leaking sensitive data is great, we may have to forego the open science principle, for instance by only allowing restricted access for reproducing the results. On the other hand, if the risk is acceptable, it may be more important to stick to the open science principles as closely as possible.

What should we do?

In sum, we need to understand that—while any personal data can pose serious risks regarding privacy—network data may pose even more serious risks because the data is *connected with each other*. Network data may be much easier to de-identify and carries a lot of revealing information about those who are in the network.

In other words, when working with network data, we need to be considerate and careful about potential harm that can be inflicted on the people in the dataset. If the

data is sensitive, we should, of course, be extremely careful about privacy leaks. But even if the data is *in principle* public, we should still think about the potential risks and potentially damaging implications of the dataset. Can someone lose their job because of this dataset and my data analysis? Can someone be publicly shamed or harmed by others because of this?

At the same time, we need to understand the broad societal context regarding social data collection and be cognizant about possible complex conflicts between individual privacy, the scientific value of research, the principles of open science, and so on. Many of these problems are difficult, having no clear answers, and call for thoughtful and nuanced approaches.

3.4 Mistakes, misconduct, and how to prevent them

Even if we address all the issues discussed above, ethical issues can still arise from the process of research and data analysis. Even with the best of intentions, researchers can always make mistakes and some can be highly damaging. And what if the researchers are the *baddies*? What if *you* are tempted to manipulate the data or exploit and distort the data to manufacture the conclusion that you want? How can we ensure that, even in the presence of bad actors, we collectively produce auditable and reproducible results free of ethical issues?

Types of misconduct

What are the types of misconduct and ethical issues that can arise during the research process? First of all, a researcher can flat out falsify data. Instead of following the normal procedures of obtaining data (e.g., surveys, web-crawling, etc.), one can potentially manufacture their own data so that it fits their preconceived conclusion. This is difficult to detect because, unlike the derived datasets where we can track its *provenance* (see Ch. 17), raw data do not have ways to check the provenance. That said, falsified data can still be detected through independent re-collection of data (and comparison) as well as discovering artifacts and anomalous patterns in the data—false data often exhibit artificial patterns. Nevertheless, falsifying the primary data is difficult to detect and can inflict lasting damages if the data gets used in many other research projects.

Perhaps worst of all, such dangers lurk at every step throughout the life cycle of data-driven research. Even if the raw data is legitimate, the processed data can still be falsified or messed up, and even with good data, faulty or misused analysis can produce erroneous results or fabrications. In fields with mostly computational research, this type of problem is easier to mitigate because others can replicate the analysis to identify problems. A more tricky issue is so-called p-hacking, sometimes known as inflation bias, which is a malpractice of trying out many different analysis (or even collecting more data) until finding a "significant" result [207]. All statistical results are suspect at best when p-hacking occurs.

And there are always errors. Throughout the process, even without any bad intentions, the researcher can be sloppy and produce erroneous results. Such mistakes can have terrible outcomes. One famous example is the so-called "Reinhart–Rogoff

error" committed by Harvard Economists Carmen Reinhart and Kenneth Rogoff, in their paper that bolstered the arguments for austerity measures in EU countries [393]. Their results were wrong simply because they failed to select some cells in their data spreadsheet [68]—affecting the lives of millions.

Why does misconduct happen?

We can think about two primary drivers: the *incentive* and the *probability of getting caught*, and if we include sloppy errors, *bad research practices*. The higher the incentive to commit a misconduct and the lower the changes of being caught, the more likely a researcher may attempt it. In academia, there is a strong incentive for publication and prestige. Decisions for hiring, promotion, and tenure are all made primarily based on publication records. Publishing in the most prestigious journals—which are hungry for surprising and strong results—can guarantee not only *survival*, but fame and funding as well. Therefore, there is a substantial incentive for misconduct—falsifying the data, fabricating results, and so on. If you can simply make up your data or force your analysis to fit your predetermined story, it becomes *much* easier to write a paper, especially one with fascinating, surprising results.

Then there is the risk of being caught. If one does not have to share the data and code, it is difficult for someone else to identify the issues. For this reason, it is increasingly common for journals to require authors to publish their data and code alongside their publication.

In industry or other organizations, the nature of the incentive may be different but it is still present. Companies may have strong reasons to exaggerate the performance of their methods to attract investments or to sell their products. At the same time, because results tends to be directly applied to real-world applications in industrial setting, it may be easier to catch any issues or errors.

Finally, let's talk about mistakes. A lot of mistakes and errors occur due to bad research practices. For instance, using software like Excel makes researchers highly prone to simple errors [356, 511]. Not performing code reviews or other poor software engineering practices are also culprits that produce errors.

What should we do?

Trusting researchers is not a solution. We need to build systems and processes that assume any researcher involved can, sadly, be a bad actor, or simply a human being who is capable of making mistakes. We have to create robust safeguards by ensuring data provenance, replicability of results, and open science. As the whole process of analysis becomes more transparent, auditable, and replicable, there will be stronger deterrence for misconduct. Fortunately, norms and policies are moving towards this direction. Increasing numbers of journals require the publication of replication data and code so that other researchers can directly replicate the results. Even when not published with the paper, sharing code and data with other researchers is becoming a strong norm across many fields. And more software engineering best practices are being embraced by researchers (see Ch. 19). While *publish-or-perish* remains a strong incentive, things are starting to improve.

3.5 Summary

Here we reviewed ethical concerns surrounding network data and network analysis. The ethical issues that we discussed often do not have clear solutions but require thoughtful approaches and understanding complex contexts and difficult circumstances. It is critical to understand how biases creeps into datasets and what kinds of negative implications they can have. It is also critical to understand the trickiness of handling private data, especially in the network context. The first step to mitigate ethical issues working with data is to be clearly aware of them.

Then we need to promote and exercise the best practices. A dataset is not complete without detailed documentation about the data collection process, potential biases, and other contexts surrounding the dataset. We must pay close attention to the bias in the data as well as privacy issues. Carefully processing the data is also important for ensuring data provenance and the correctness of the analysis. Embrace best practices in software engineering to minimize errors in computerized analysis. Ensure replicability of the data processing and results.

Bibliographic remarks

All scientists should be ethical and responsible in how they conduct their research. For general introductions to ethical and responsible research, we recommend Resnik [394], Seebauer and Barry [422] and, especially for the social sciences, Israel and Hay [228].

Our increasingly data-rich, digital world creates new perils. For an engaging, general audience treatment of how data and data algorithms lead to negative societal outcomes, we highly recommend O'Neil [352].

Network data in particular introduce special concerns. Staiano et al. [443] explore ways to infer personality profiles of users through their social network data. Garcia [173] discuss how online platforms can build latent or "shadow" profiles of their users. Sarigol et al. [414] describe how privacy in a networked world ceases to be an individual choice but instead depends on the group. Bagrow et al. [30] demonstrate how actionable, predictive information can be extracted about individuals using only their social ties. The privacy implications of network data remain an active area of research.

Exercises

3.1 (**Focal network**) Consider the Malawi Sociometer Network (Sec. 2.4). The data underlying this network were gathered as part of an experiment where participants wore small devices that could detect who they were in contact with. This is an example of a *human subjects* research project.

(a) Read the primary publication describing this dataset [353]. Summarize their discussion of the study ethics.

(b) Are there other ethical concerns? Were you to conduct a similar study, what ethical concerns might you need to address?

https://doi.org/10.1017/9781009212601.005 Published online by Cambridge University Press

3.2 A company wishes to hire you as chief data scientist. They seek to distribute a free, location-based smartphone game where players gain points by exploring their surroundings, promoting exercise and enjoying the outdoors. Their business model is to collect data from the phone sensors, such as Bluetooth and wireless Internet, allowing them to track users' whereabouts and what other devices connect with their phones, then resell this data to businesses seeking advertising profiles, shopper analytics, and more.

(a) What ethical concerns are posed by this company's plan?

(b) What options might the company have to address ethical concerns?

3.3 (**Advanced**) A social media provider falls victim to a cyberattack. Hackers leak user details, including email addresses and the usernames of their online friends/followers.

(a) Describe some ways that users can be negatively affected by revealing this information.

(b) Suppose at some earlier point the social media provider asked users to provide access to their address books, giving them a list of known contacts. Not all these contacts are users of the platform, giving the platform a new source of potential users. If *those* details are also found by the hackers, how might non-users be negatively affected by the leak?

(c) Suppose the leak included information describing what devices were used by users to access their account. This information may be cross-referenced, allowing someone to "join together" a user's multiple accounts. Those users may have very important reasons for those separate accounts. Describe some ways users can be negatively affected if it becomes public knowledge that they have multiple accounts. How serious could this be compared to other concerns?