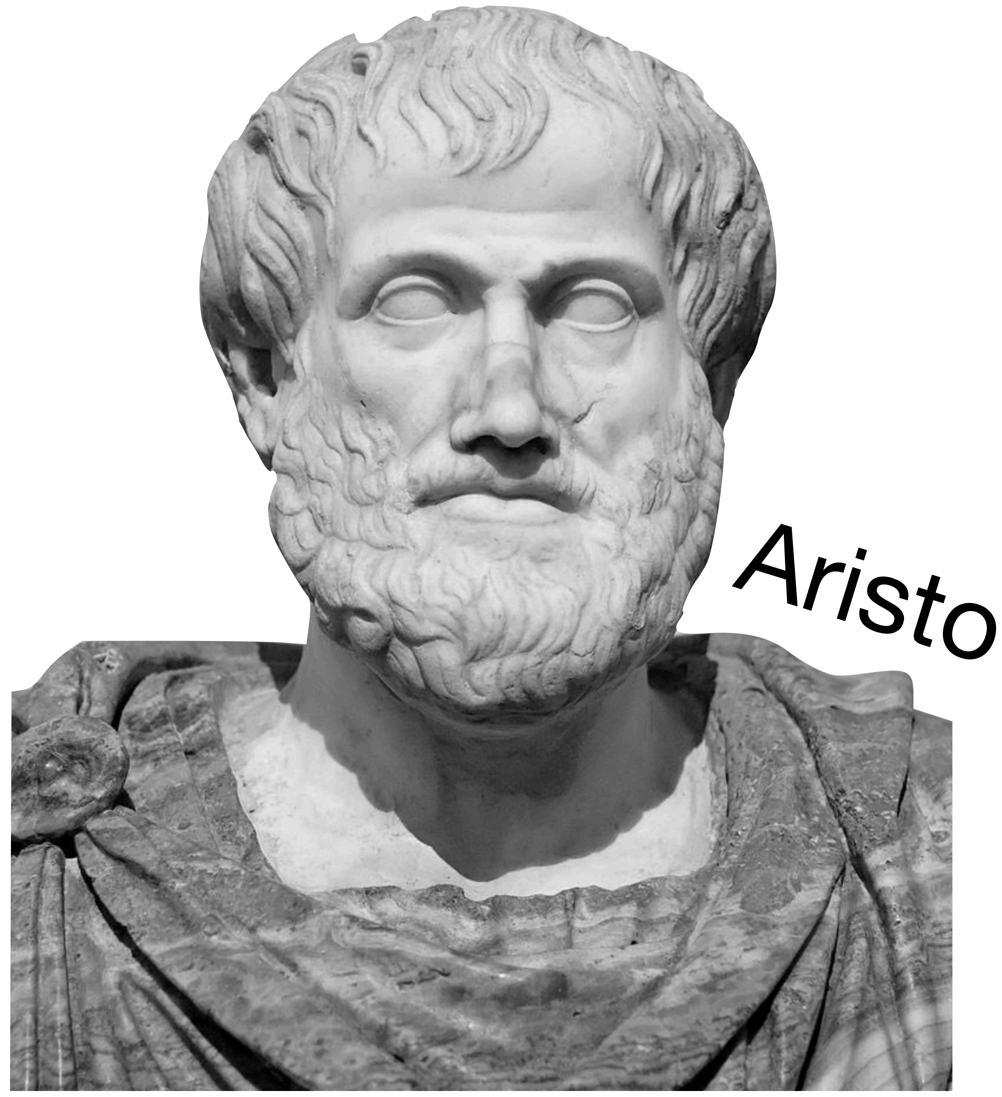
Research Design (also, How to read a research paper)

Introduction to the scientific (publication) method

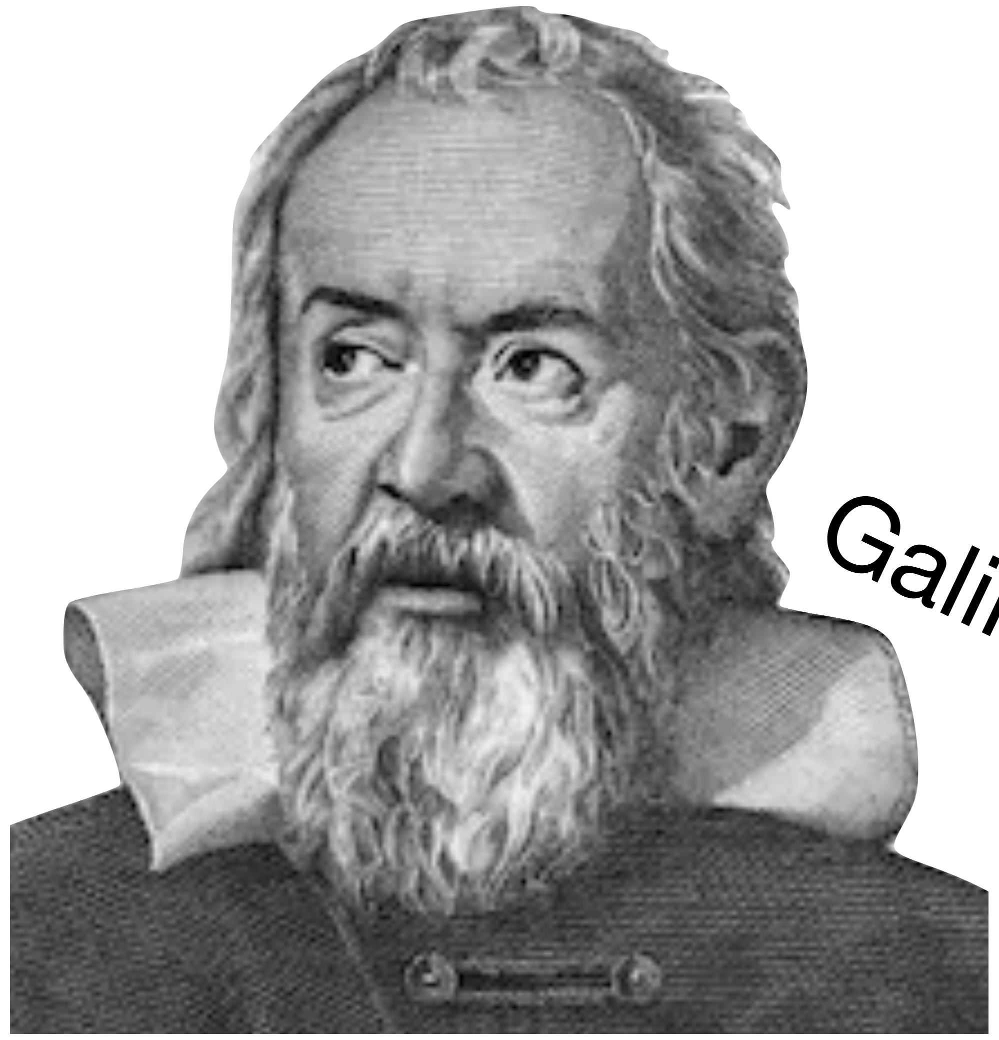


Aristotle

Knowledge from the rational analysis of nature

Ibn al-Haytham

Experiments and reproducibility

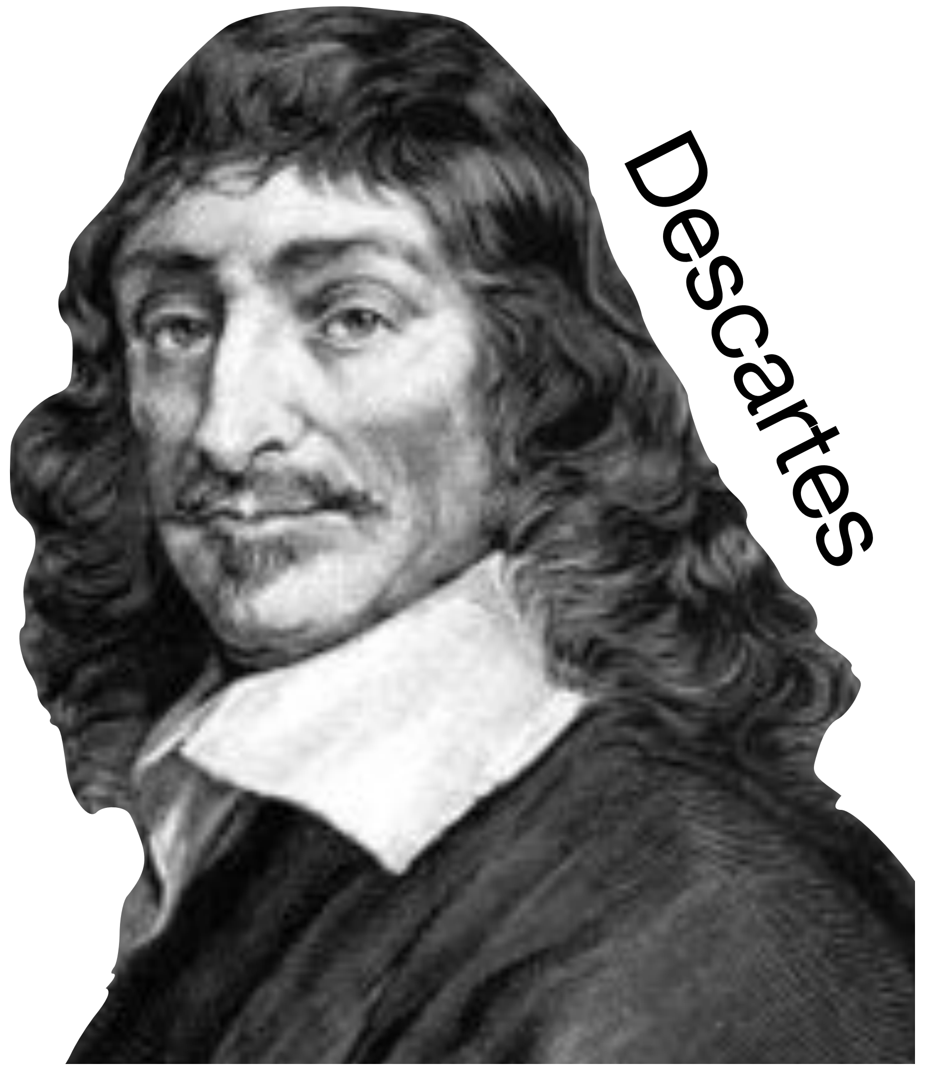


Galilei

Independent, rational Confirmation

Bacon

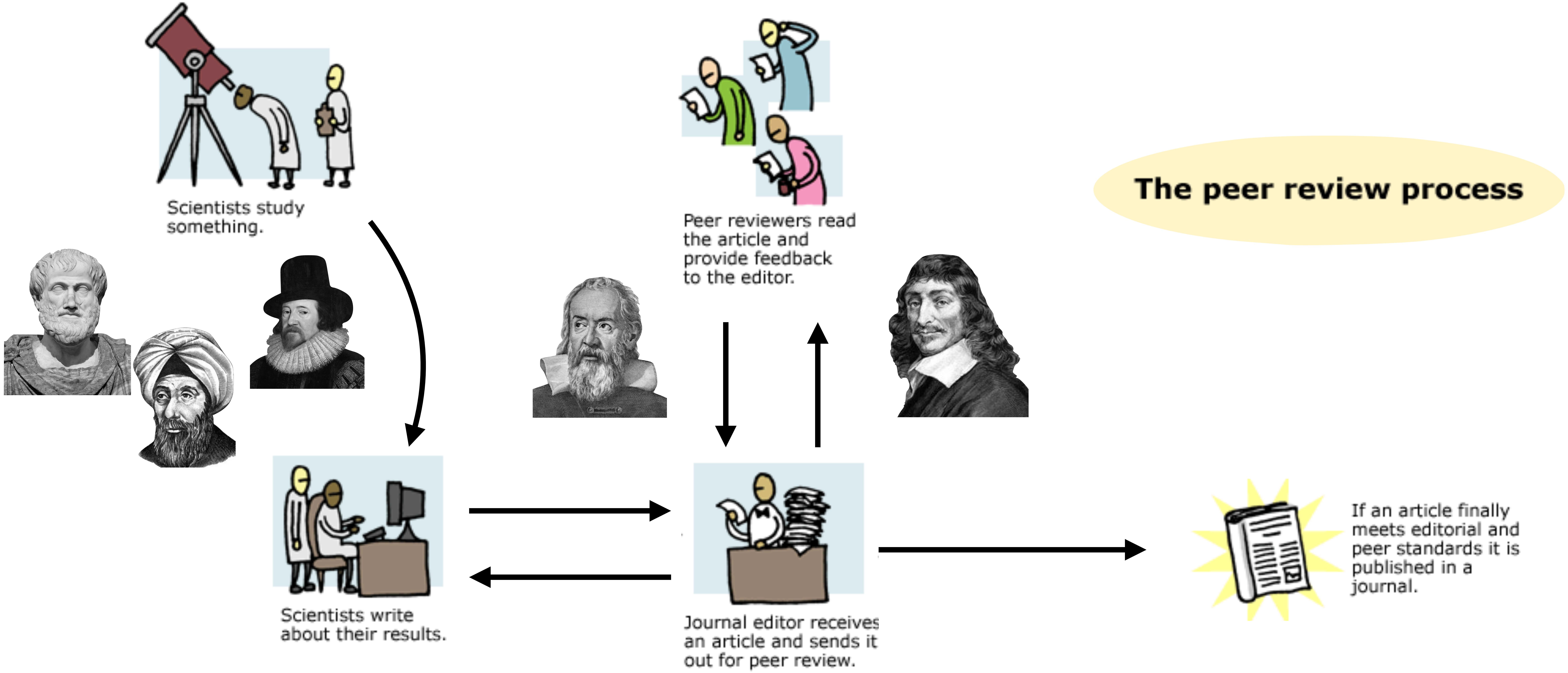
First-hand Experience



Descartes

How do we know what we know? (Math can help)

Introduction to the scientific (publication) method



Scientists study something.

Scientists write about their results.

Journal editor receives an article and sends it out for peer review.

Peer reviewers read the article and provide feedback to the editor.

If an article finally meets editorial and peer standards it is published in a journal.

The peer review process

Elements of a Research Paper: Context

What is the general context of the paper?

The context include:

- the general field (e.g., literature, history, archeology, tourism, biology, forensics, religious studies);
- the specific application (e.g., literary analysis, quantitative history, genetics, virology, forensics intelligence, tourism planning, biblical quantitative studies).

Elements of a Research Paper: Problem/Motivation

- What are the problems the authors want to address?

- Why are those problems important (impact, theoretical and/or practical needs, etc.)?

- What are the main contributions of the paper?

Elements of a Research Paper: Data

- How did the authors gather their data?
- Did they digitise their data? How? Is the material publicly available?
- What tools did they use 1) to handle (store, manipulate) the data and 2) to compute measures on the data?
- What measures did they apply?

Elements of a Research Paper: Results

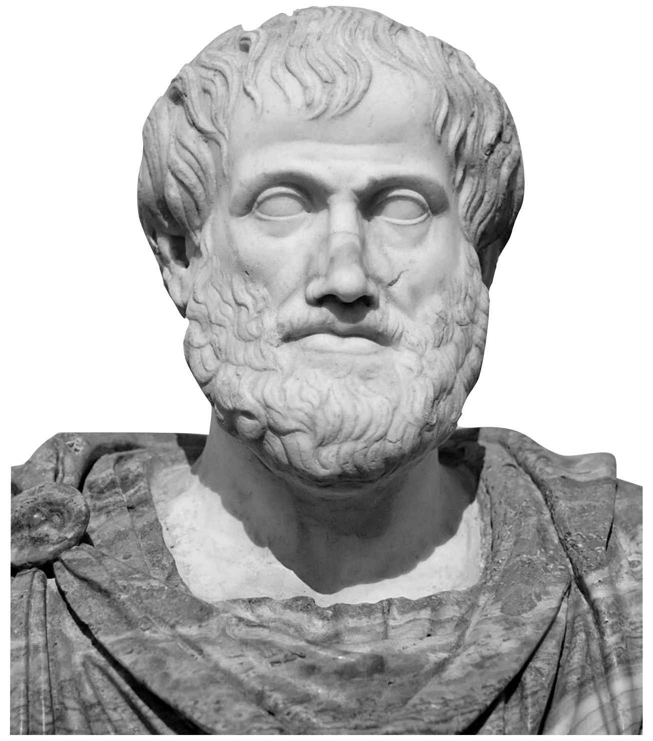
- What is the connection among:
  - the gathered data;
  - the applied measures;
  - the properties found.

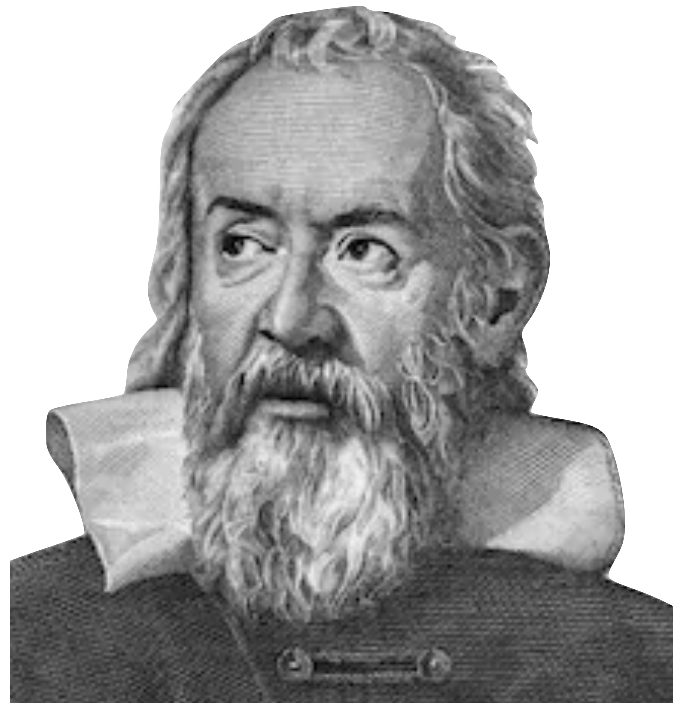
Peer-reviewing: Critique

- What is your opinion/critique on the paper?
- Do you think the contributions solve the problems they presented?
  - To which extent (completely, what parts)?
  - Why?
- What could the authors have done differently to answer their research problems (e.g., gather data with additional information, build their model differently, apply alternative measures)?

Elements of Network Analysis Research

There is a common denominator to all network analysis research: the usage of a tool from discrete mathematics, called **graph theory**, to *reduce and draw conclusions from naturally-occurring (network) phenomena*





Whole-network and Personal-network Research Design

Two fundamental kinds of network research designs:

Whole network: we study the set of ties among all pairs of nodes in a given set, e.g., we study who is friends with whom among all members of a given organisation - there, the relation being measured is a dyadic variable that has a value for every pair of nodes (every dyad might be assigned a 1 or a 0, whether they are friends or not).
Whole-network designs enable researchers to employ the full set of network concepts and techniques, which often assume that the entire network is available. However, the cost of assembling and managing the network can quickly rise due to the whole-network scope.

Personal network: we study a set of nodes called "egos" and their ties to others, called "alters" (not necessarily among the set of egos). Personal-network designs have the advantage of simplifying the gathering and management of the network.

Sources of Network Data

Primary sources: the researcher collects the data first-hand, e.g., by asking questions to or observing the behaviour of the interviewees, from a survey, probing a (computer) network or applying other kinds of first-hand measures on the focussed network.

Secondary sources: the researcher gathers data that already exists somewhere, whether in paper records (e.g., fish records, historical marriage records), or electronic databases (e.g., emails, social networking sites). Secondary data is often easier and quicker to collect but imposes strong and arbitrary limits on the type of relations studied. Some of the computer-based data generated by social media such as Facebook and even email represents a transitional form between primary and secondary data. Although the data is collected directly, as in primary research, there are limitations on the types of relations available for study, as in secondary research.

Types of nodes and types of edges/ties

Relational states							Relational events	
Similarities			Relational roles		Relational cognition			
Location	**Participation**	**Attribute**	**Kinship**	**Other roles**	**Affective**	**Perceptual**	**Interactions**	**Flows**
Same spatial and temporal space	Same clubs, same events	Same gender, same attitude	Mother of, sibling of	Friend of, boss of, student of, competitor	Likes, hates	Knows, knows of, sees as happy	Sold to, talked to, helped, fought with	Information, beliefs, money

Types of nodes and types of edges/ties

Category	Varieties and examples
Co-occurrences	Co-membership in groups Co-participation in events Physical distances Similarities in attributes (e.g., political view)
Social relations	Kinship relations Affective relations (e.g., dislikes) Perceptual relations (e.g., knows)
Interactions	Transactions (e.g., ‘sells to’) Activities (e.g., ‘sleeps with’)
Flows	Ideas and information Goods Infections

Types of nodes and types of edges/ties

Co-occurrences: it is relatively easy to collect, e.g., membership type data is often not thought of as particularly private or sensitive. It is often available via archival sources (e.g., people serving on the boards of directors of firms, IMDb for the cast or crew of films).

Social relations: ties of continuous nature and they are relational states. Many social relations are institutionalised (e.g., being married) rather than being a perception of the individuals, and can be collected from sources other than the people directly involved (e.g., community, family members, archival records). Other types of social relations have no independent existence, such as affective and perceptual ties, and such has to be obtained by surveying the perceiver.

Types of nodes and types of edges/ties

Interactions (1/2): can be either directly observed or reported on by respondents. Although interactions are often directly observable, there are *always issues of interpretation*.

In a network study of a fish camp, Johnson and Miller (1983) observed two Italian fishers engaged in what appeared to be a heated discussion. Johnson asked a younger Italian fisher, who was also observing, what the conflict was all about. The younger Italian explained that there was no conflict, but that the two men – who were brothers – were simply having a friendly discussion about a nephew. Johnson was interpreting that interaction from his cultural perspective rather than from the perspective of the two Italians engaged in the interaction.

Types of nodes and types of edges/ties

Interactions (2/2): The higher the level of interpretation, the more theoretically useful the data is likely to be (e.g., passing from a “friendly conversation” or a “heated argument” to “had a face-to-face interaction”) at the detrimental of the specificity of the study.

Frequently interactions are collected as a *proxy for unseen underlying social relations*, e.g., who-talks-to-whom outside of work **might** be a proxy to mean the speakers are friends. The same can happen when acquiring data through archived text exchanges (e.g., chat messages, emails), where it is not always clear what is the kind of relationship that ties the two writers.

Types of nodes and types of edges/ties

Flows: can be seen as the outcome of interactions (e.g., people exchanging information as well as sharing-networks among subsistence hunters). Flows data is difficult to obtain at the individual level (the amount of information obtained from a conversation), while organisation-level flows data is easier to obtain, thanks to records and accounts, e.g., the dollar-value of flows of raw materials and manufactured goods between countries, personnel flows between companies, universities or football teams.

Sampling and bounding

When is enough (data) enough?

More specifically (to network analysis) what is the bound of the set of nodes to be included in a study?

The problem is subtle and ***it does not really concern the size of the network*** but rather the nature of the research question. E.g., if we want to determine what are the main actors determining the choices into a company, what is the scope we consider? Just the company? Should we consider the families of the employees? What about their friends and clubs? Other employees in competing companies?

Sampling and bounding, guidelines

If the research cannot restrict the set of alters that a respondent could name, use a personal-network research design. In a sense, the boundary specification problem involves two sets of actors that need bounding: the egos (in whose ties we are interested), and the alters (those with whom egos have ties).

Consider whether the object of the study is a sociological group or not. Groups are recognised by their members and they have boundaries: part of the concept of a group is that there are members and non-members, even if in fact the boundaries are fuzzy and/or contested. If one is studying the internal network of a group, then getting the boundaries more or less right is important. Also artificially-defined (by the researcher) groups do not necessarily threaten the validity of the research design, as long as the grouping choice is rationally defined/explained by the research question.

Strategies for finding group boundaries

Most groups have fuzzy boundaries.
Even formal groups such as corporations, can have fuzzy boundaries, e.g., does a corporation include part-timers? Sub-contractors? New hires? Applicants? Retirees? Consultants?

Snowballing: sampled individuals recruit new subjects from among their acquaintances. This sampling technique is often used in hidden populations, which are difficult for researchers to access. However, snowball samples are subject to numerous biases (e.g., people with many ‘ties’ are more likely to be recruited into the sample).
If the purpose of the study is to *discover the nature of ties* that connect the areas of high redundancy or density in social networks, then ties bridging these areas of high density need to be pursued and a *redundancy criteria* may need to be applied across several waves of a snowball sample.

Strategies for finding group boundaries

Nominalist and realist criteria: respectively also called “etic” or “emic” criteria to identify the members of the studied group. The term “etic” indicates a universal (nominalist) objective, third-person criteria to classify the individuals. The term “emic” indicates the recognition (realist) of belonging to a given group either by members of the group itself or by a set of ideologies and behaviours recognised as characteristics by the members of that group.

For example, in a study on innovations through a network of commercial fishers, to define the boundaries between professional and hobbyist fishers, the researchers used the fishers’ own perceptions (emic) to refine the sample, asking all licensed fishers (etic) to identify who where those other fishers that they considered full-time professionals.

Sources of data, reliability and validity issues

Validity: how closely the model we have represents reality? Otherwise, “ex falso, quodlibet”.

Sources of data, reliability and validity issues

Validity: how closely the model we have represents reality? Otherwise, “ex falso, quodlibet”. Validity errors include:

- **Omission errors:** missing edges and nodes have huge impacts on errors in network variables (e.g., in centrality measures), by making the network appear more/less disconnected than it really is or make nodes and edges in the network appear to be more “important” than they really are.

- **Commission errors:** Like omission errors, the erroneous inclusion of nodes and edges can affect the ultimate determination of node-level measures and the identification of key nodes.

Sources of data, reliability and validity issues

Validity: how closely the model we have represents reality? Otherwise, “ex falso, quodlibet”. Validity errors include:

- **Edge/node attribution errors:** mis-assignment of a behaviour to a node can yield attributed linkages in a network that in reality do not exist. Attribution error is a common problem in the interpretation of two-mode data that has been converted to one-mode. For example, two individuals in a university program may co-attend a large number of classes. We therefore assume a connection (either pre-existing or as a result of meeting in class). But it could easily be that one of the individuals is a non-traditional student who is older and married and does not hang out with other students in the program. Treating co-attendance as a tie is, in this case, a mistake. Collection of other relational data could help in determining whether an active tie actually exists in this case.

Sources of data, reliability and validity issues

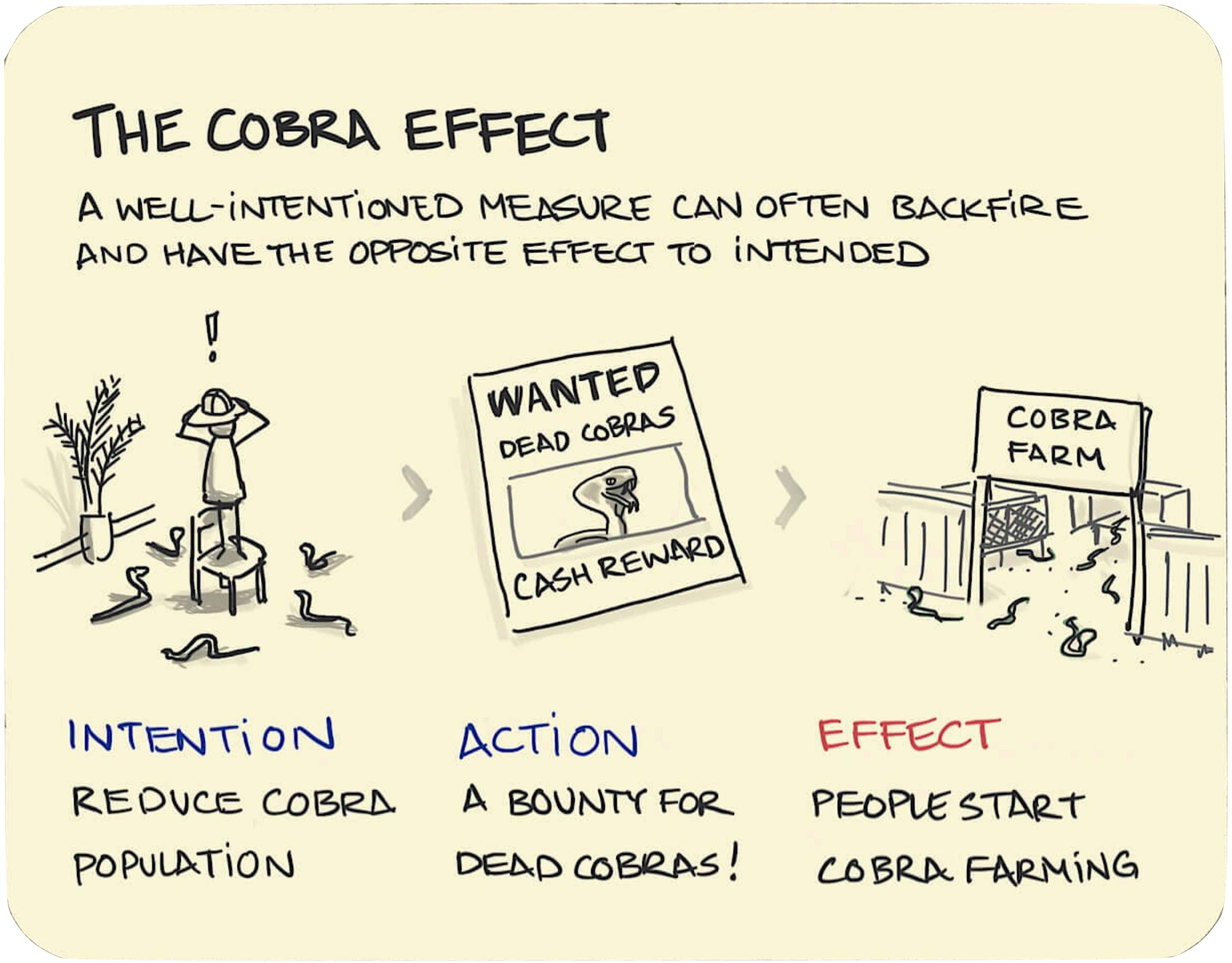
Validity: how closely the model we have represents reality? Otherwise, “ex falso, quodlibet”. Validity errors include:

- **Data collection and retrospective errors:** care should be taken when using network data collected from individuals where the network elicitation question deals with reports of behaviour, particularly having to do with social interactions of a temporally discrete nature. Avoid generic questions and cross-check (if possible) assertions by your sample, as people, when describing their behaviours (e.g., "whom did they interact with yesterday?") can both make commission and omission errors.

Sources of data, reliability and validity issues

Validity: how closely the model we have represents reality? Otherwise, “ex falso, quodlibet”.

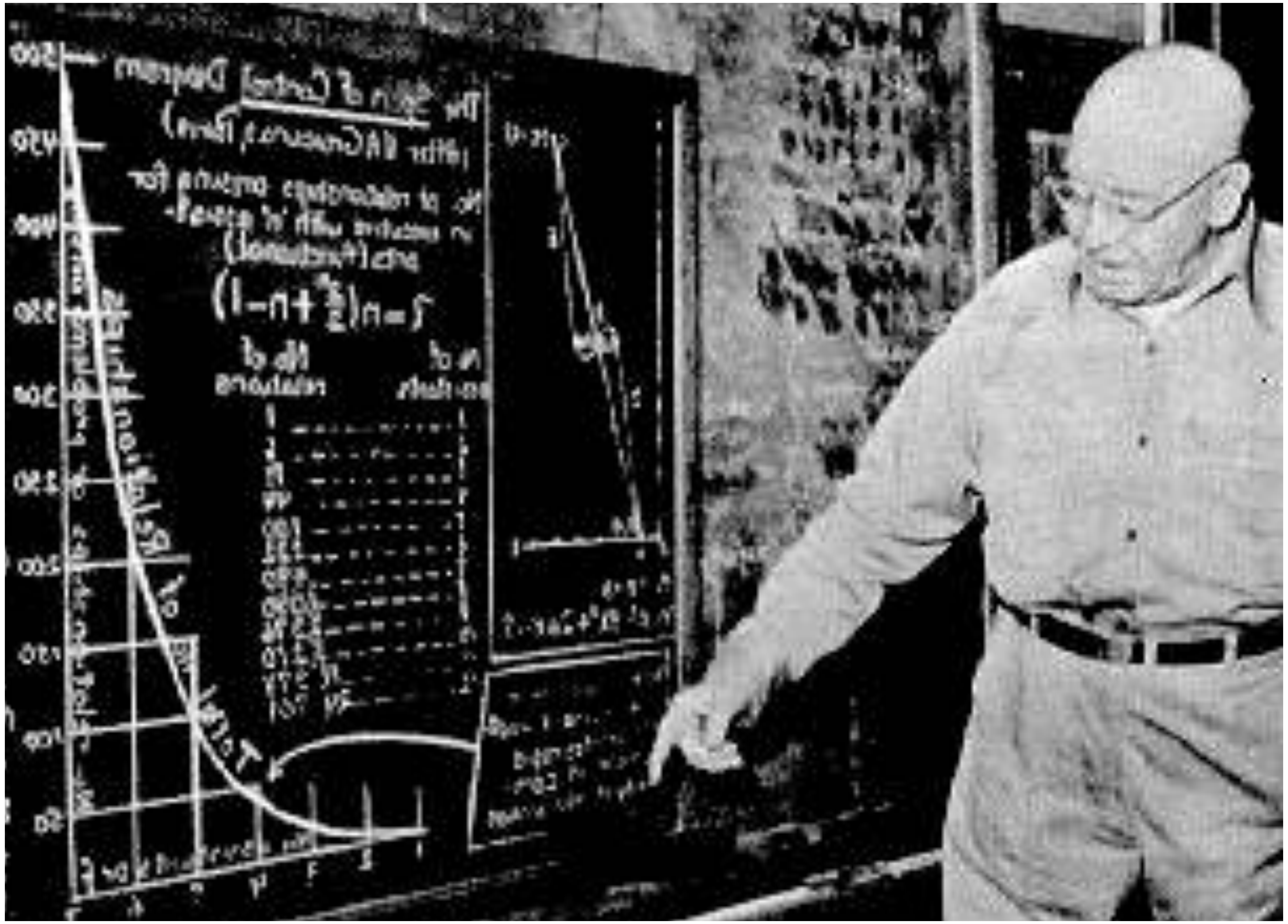
Timeo proxies et clausolas ferentes


THE COBRA EFFECT
A WELL-INTENTIONED MEASURE CAN OFTEN BACKFIRE AND HAVE THE OPPOSITE EFFECT TO INTENDED
!
WANTED
DEAD COBRAS
CASH REWARD
COBRA FARM
INTENTION
REDUCE COBRA POPULATION
ACTION
A BOUNTY FOR DEAD COBRAS!
EFFECT
PEOPLE START COBRA FARMING

Sources of data, reliability and validity issues

Validity: how closely the model we have represents reality? Otherwise, “ex falso, quodlibet”.

A **map *is not* the territory** it represents, but, **<u>if correct</u>, it has a *similar structure*** to the territory, which accounts for its usefulness.



A Non-Aristotelian System and its Necessity for Rigour in Mathematics and Physics, Korzybski 1931

Sources of data, reliability and validity issues

Reliability: how consistent is the model we assembled (e.g., a standard scale to measure people's height?). If we do not apply a general measure but an ad-hoc one, our results are not reliable (to check and use by others). Threats to reliability include:

- **Data fusion/aggregation:** when aggregating data on different temporal, relational and spatial scales, it is possible that important individual nodes and edges are excluded because they have lost their importance in the network. Thus, there should be some guiding principles - preferably of a theoretical nature - for making aggregation decisions (e.g., before and after a hypothesised important event).

Sources of data, reliability and validity issues

Reliability: how consistent is the model we assembled (e.g., a standard scale to measure people's height?). If we do not apply a general measure but an ad-hoc one, our results are not reliable (to check and use by others). Threats to reliability include:

- **Errors in secondary sources and data mining:** secondary-source data may have inherent biases, which should be considered in any analysis. Second-source data might be easier to collect than primary one, but it can be fraught with errors at a variety of levels. When adopting secondary data it is important to question the consistency of the model, asking questions like "if we asked a survey question, what survey question would the tie(s) in the model correspond to?"

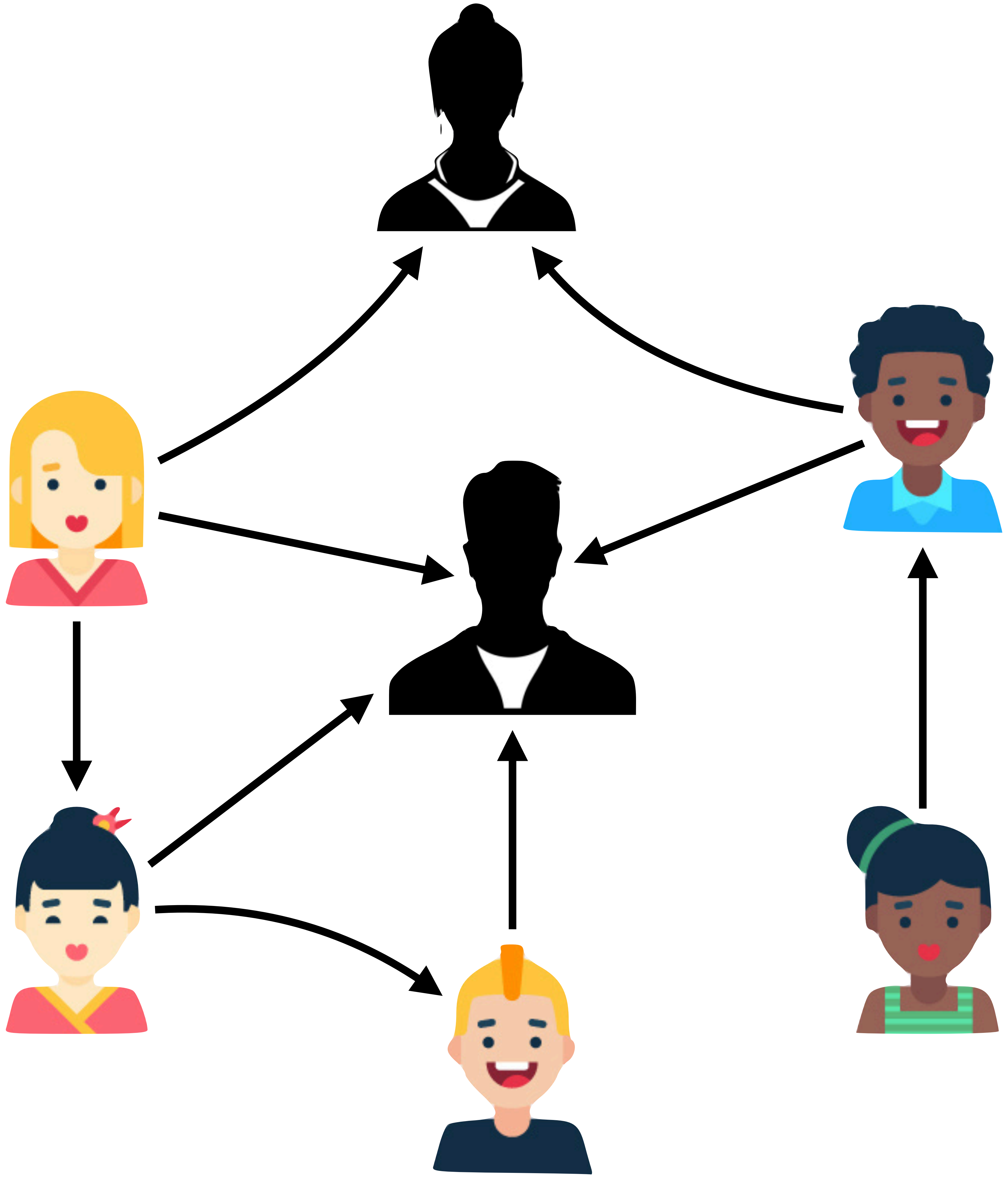
Sources of data, reliability and validity issues

Reliability: how consistent is the model we assembled (e.g., a standard scale to measure people's height?). If we do not apply a general measure but an ad-hoc one, our results are not reliable (to check and use by others). Threats to reliability include:

- **Formatting errors:** when mining data (or the Web) errors can derive from (unexpected) differences in document formatting. These errors can lead to the over- or under-representation of terms, actors, attributes, etc. in the data retrieval process. Care should be taken that any relations assigned among nodes is not an artefact of formatting errors. In addition, Web scraping and automated data mining methods should be scrutinised for consistency of study-important concepts. The bottom line is that the quality of a study is a function of the quality of the data: garbage in, garbage out.

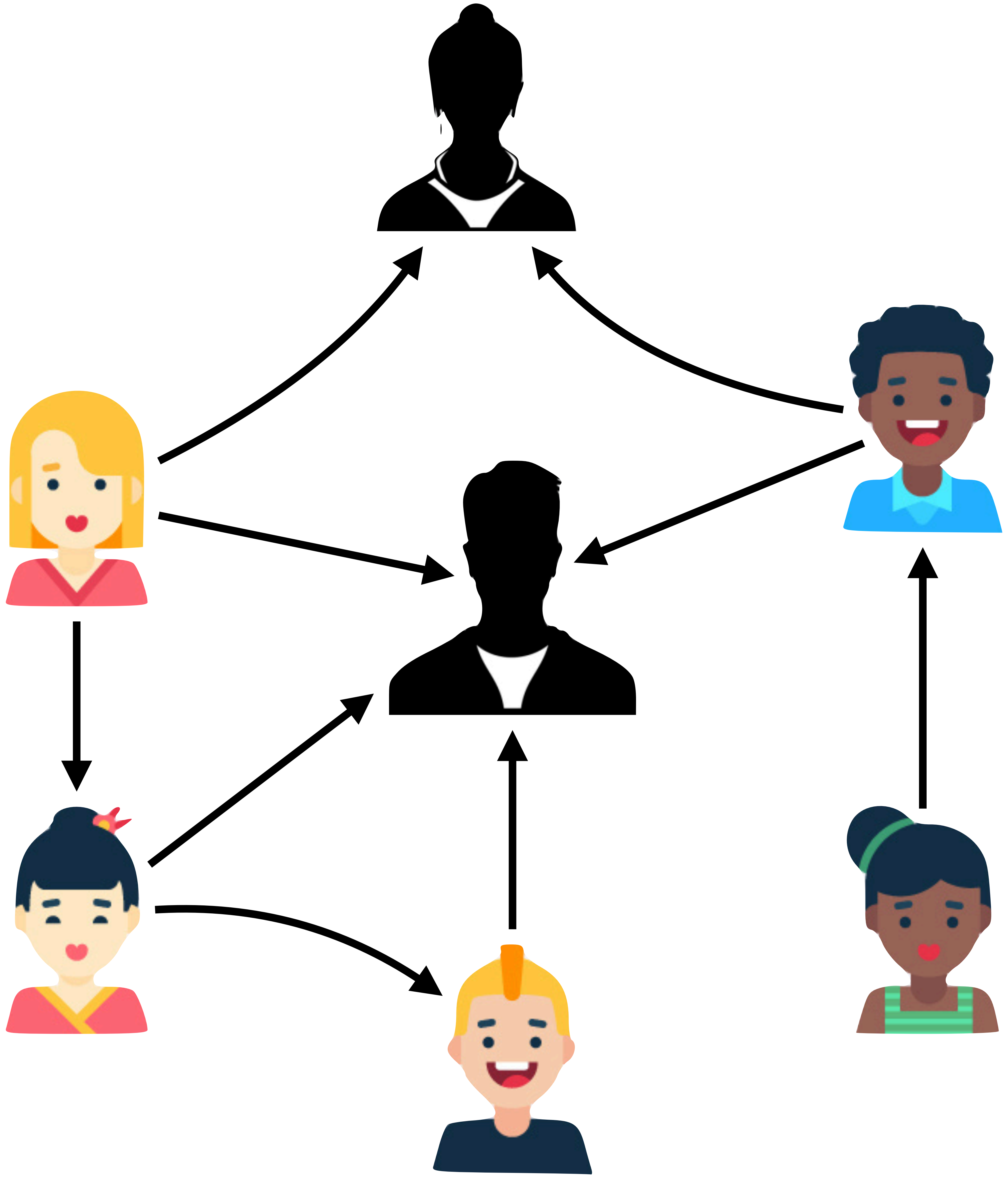
Ethical Considerations

A key data-collection issue of network studies is the impossibility of anonymity: the **respondent must identify themselves**, which means **the researcher can only offer confidentiality**. This makes it imperative to make it clear to the respondent who will see the raw data and what can reasonably be predicted to happen to the respondent as a result of an accidental breach of confidentiality.



Ethical Considerations

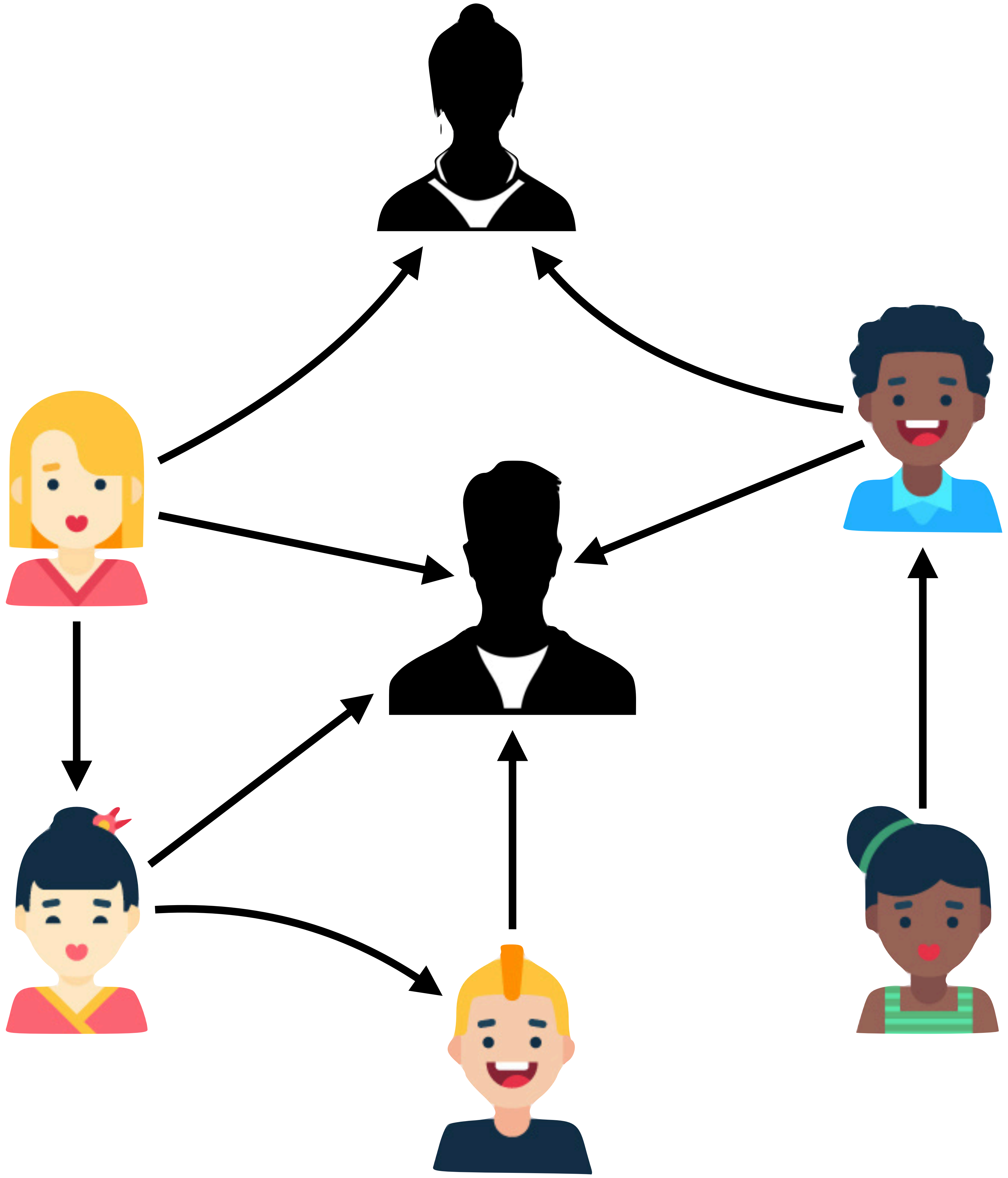
Corollary. **non-participation by a respondent in a network study does not necessarily mean that they are not included in the study**, as other respondents may still list that person as a friend, enemy, etc. E.g., a person who does not wish to be embarrassed by their poor standing in the group will still be found to be the person most often named as difficult to work with.



Ethical Considerations

We can eliminate all explicit non-respondents from the dataset altogether, however this may lower the quality and representativeness of the data.

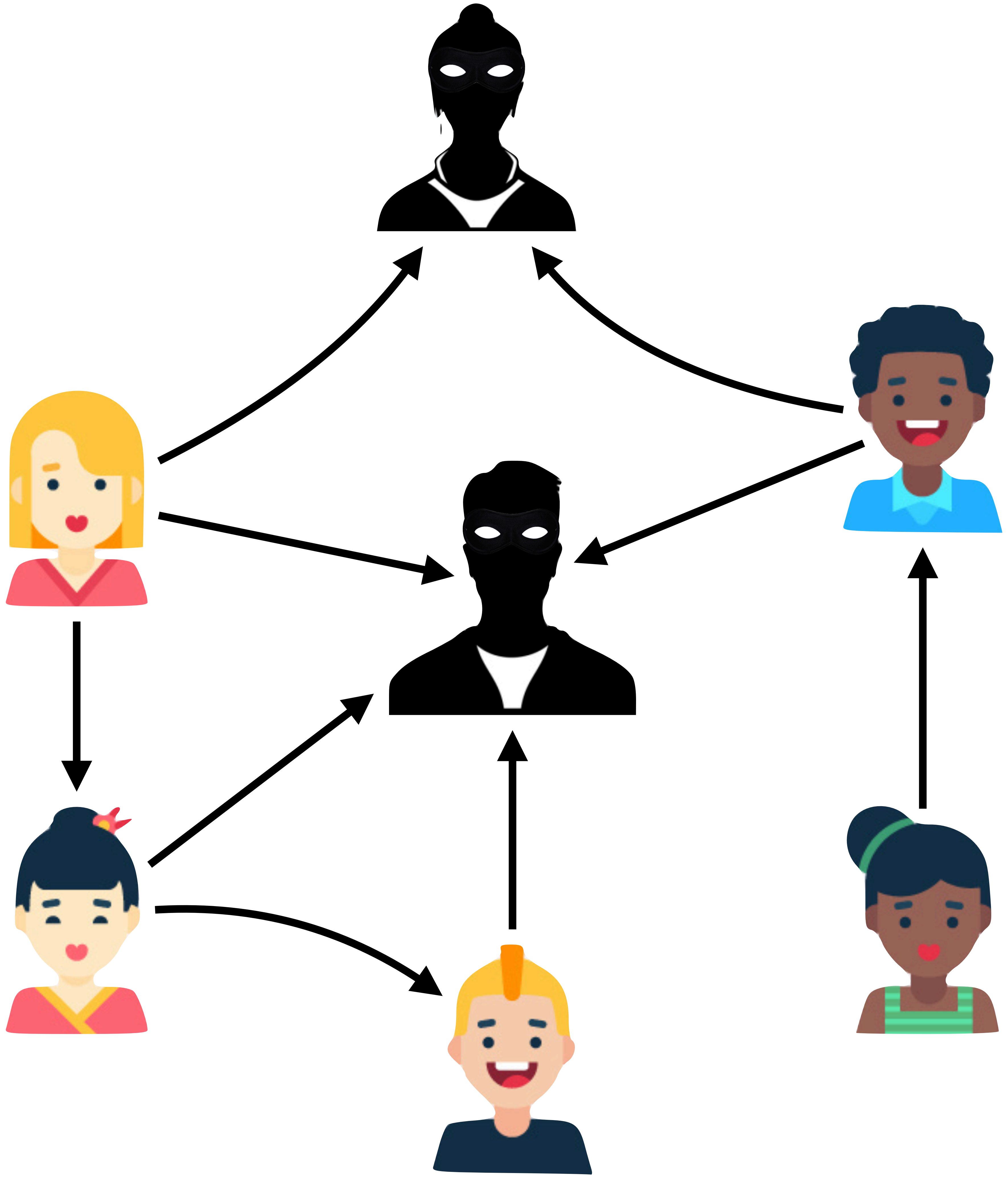
This is particularly a evident in applied settings, where decisions are based on the results of the study. The researcher can warn the management of the problem, but it is likely they do not have the sensibility to appreciate the depth of the problem - they might also need the information that the researcher is trying to suppress to understand the problem.



Ethical Considerations

This issue is at the foundation of network studies, as respondents report on their relationships with other people, some of whom may not wish to be reported on.

While the matter does not seem unethical per-se - the respondents are free to own and report her own perceptions - e.g., when dealing with illegal activities, there is a clear implication that the named party does in fact do **illegal actions**, from stealing to sexual harassment



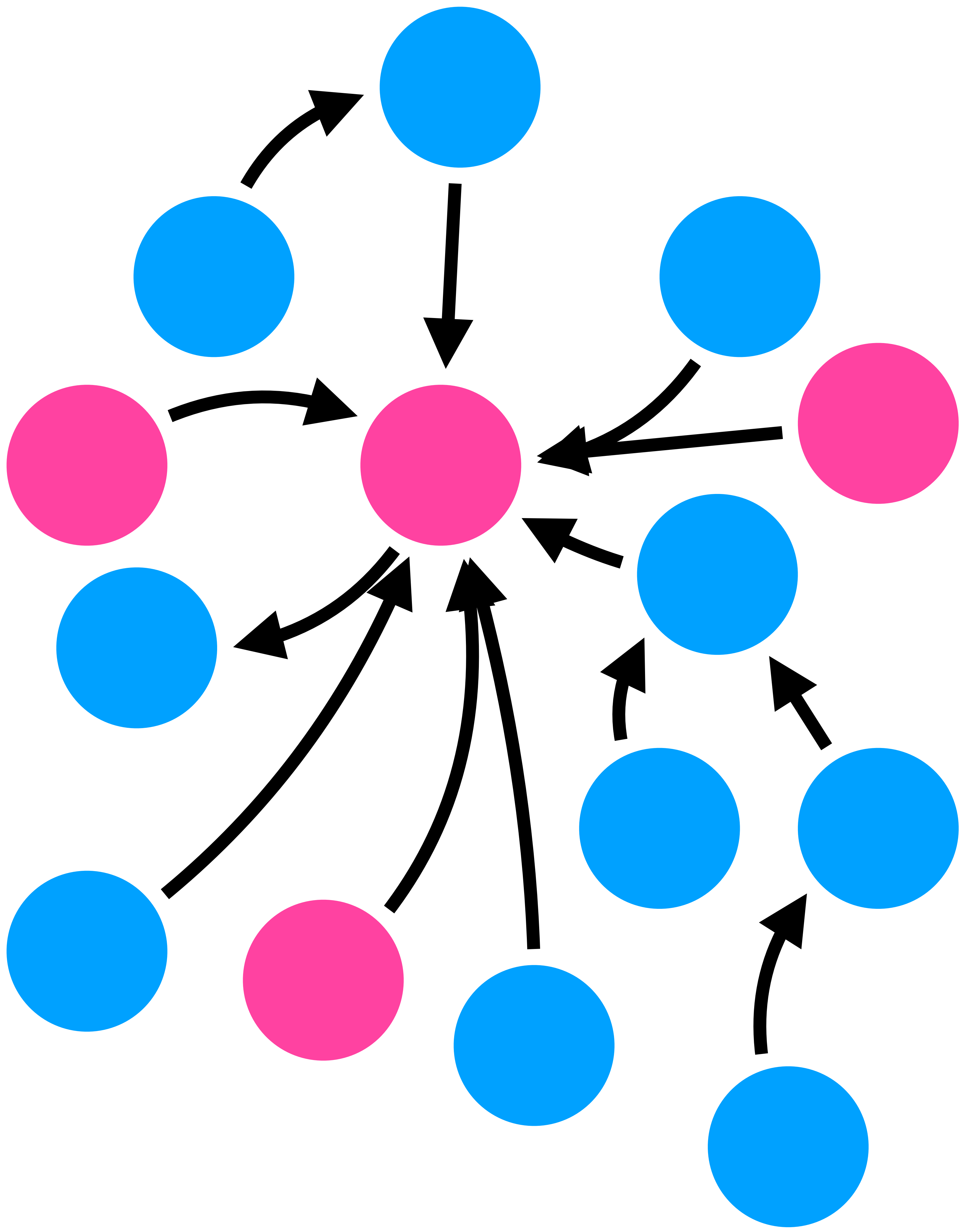
Ethical Considerations

In network studies missing data is exceptionally troublesome, e.g., when a few highly-central players are missing, the resulting network could be sensibly different than if those people responded.

This creates incentives for researchers to **“press” respondents to participate**, underestimating their risks and “coercing” them (e.g., though management’s pressure) to unwillingly participate, possibly tainting their mood/feelings towards the study and possibly their responses.

Ethical Considerations

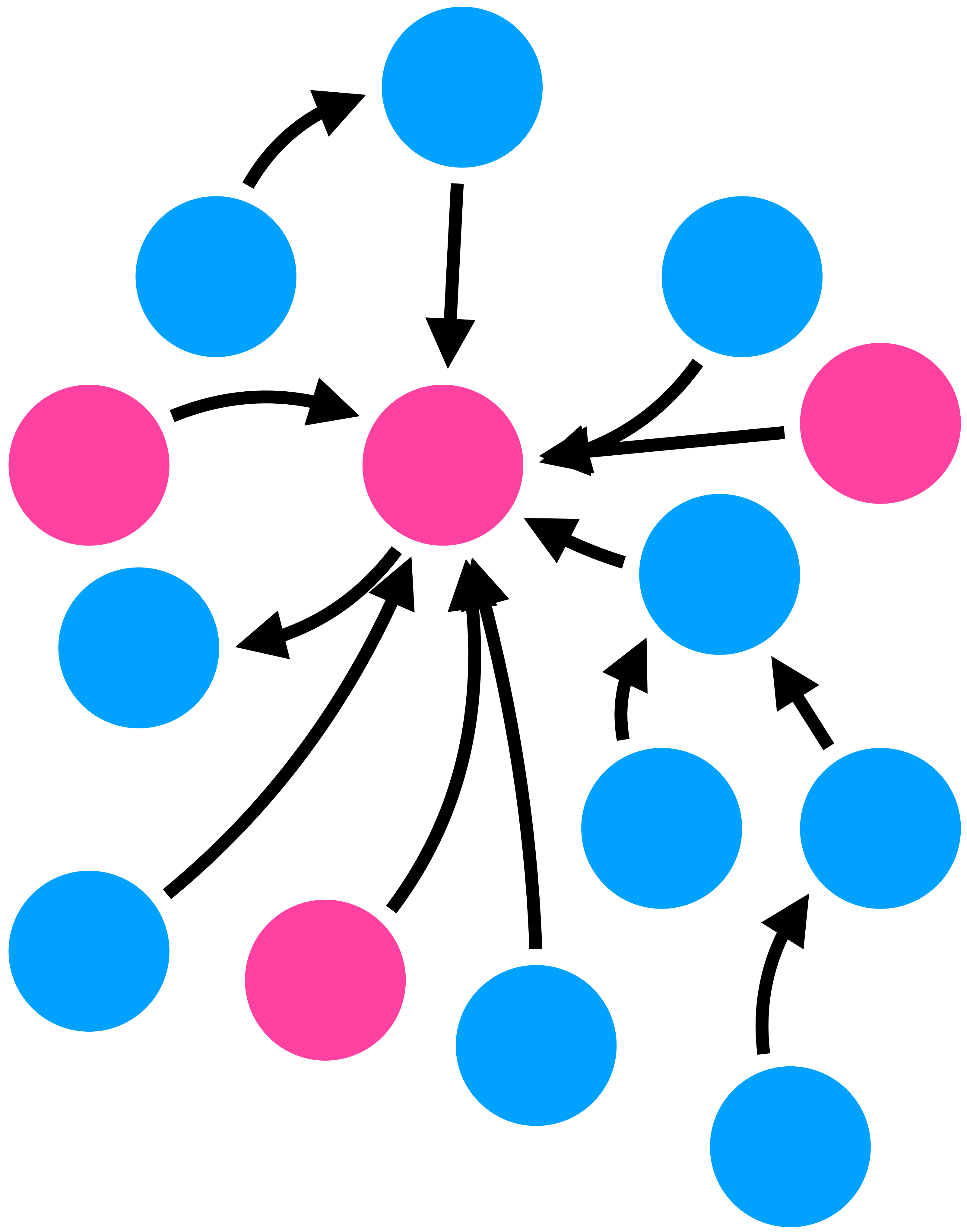
In network analysis it is common to present a network diagram that shows who is connected to whom. When doing so, using a low-level display with un-labelled nodes that represent the raw data might not be enough to ensure the responders' privacy: e.g., organisational **members can deduce the identity of one person** (like the only high-ranking woman in the Boston office) and once that person has been identified, their known associates can sometimes be deduced as well, eventually unraveling the whole network.



Ethical Considerations

Most respondents have neither participated in a network study nor have an intuitive understanding of the ramifications of their responses. This is true also for the management, that could need guidance in understanding the subtleties of the conclusions from the study.

This puts an **ethical burden on the researchers to be clear and provide enough background about the risks and implications of their study**.



Let’s read and analyse our first scientific paper