UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Confidentiality

26–28 September 2023, Wiesbaden

The risk of identity disclosure through network structure: anecdotal evidence from a hackathon

M.M. de Vries[1], R.G. de Jong[1,2], M.P.J. van der Loo[1,2], P.-P. de Wolf[1], F.W. Takes[2]

[1]Statistics Netherlands
[2]Leiden University

mm.devries@cbs.nl

Abstract

The probability of disclosure is determined by two factors: the probability of disclosure conditional on a certain scenario of attack and the probability of that scenario taking place. Most statistical disclosure control policies assume a worst-case scenario by setting the second factor equal to one. In the case of new types of data sets, it is worth investigating this assumption.

Statistics Netherlands has recently developed population-scale network data where nodes are persons and links represent various real-world connections including family, household, work, school, and geographical connections van der Laan et al. (2022). In this context, we have developed an anonymity measure where it is assumed that an attacker has certain prior knowledge about the network structure surrounding a node de Jong et al. (2023a).

To gain insight into how likely it is that an attacker obtains such knowledge, a hackathon was organized where students were challenged to discover real-world connections surrounding a selection of persons who volunteered to participate. In a time of about four hours, a group of 22 students found more than 5,000 typed links surrounding 26 volunteers by searching or scraping the web. Students were asked to judge the reliability the of link and link type and register the source of information. The results of the hackathon were partly checked by the volunteers.

Analysis of this data set provides anecdotal evidence for differences in the ease with which different link types can be found online. Although perceived relatively unreliable, social media ('friend') links are relatively easy to obtain while links related to geographical vicinity and household sharing appear difficult to find. We also find differences between reliability estimates by the hackathon participants and the reliability indicated by the volunteers, depending on link type and source of information.

1 Introduction

The arrival of the world wide web, and especially that of online social platforms, has created a new challenge in the field of privacy. Scandals such as that involving Cambridge Analytica in 2018 Confessore (2018) have increased public awareness of privacy issues with social media, but nevertheless the percentage of Dutch citizens that participate in online social networks (OSN's) keeps increasing Centraal Bureau voor de Statistiek (2022). Technologies such as Open Source Intelligence (OSINT) already take advantage of the wide array of information available online, but most traditional statistical disclosure methods do not account for this availability de Vries et al. (2021).

Statistics Netherlands has recently developed a population-scale network of the Dutch population, where all Dutch inhabitants are included as nodes van der Laan et al. (2022). For every citizen, the edges represent links between family, colleagues, household members, neighbours or schoolmates. Networks of this type and size are new for Statistics Netherlands and although useful for getting an understanding of population-scale network phenomena Bokányi et al. (2023), they bring along new challenges for statistical disclosure control. An anonymity measure has been developed that quantifies risk on the assumption of certain prior knowledge on the side of the attacker de Jong et al. (2023a,b). To be able to make a better assessment of the true risks linked with these types of networks, we are interested in exactly how likely this prior knowledge is.

Specifically, we are interested in the ease with which one can determine a person's social circle using online sources. While research has been done to outline what personal information is available, either on the broader web Khanna et al. (2016); Pastor-Galindo et al. (2020) or specifically on OSN's Aliprandi et al. (2014); Koot et al. (2014), it is harder to find sources that detail the ease with which certain personal information can be retrieved online.

We therefore set out to organise a hackathon, where we encourage participants to reveal exactly which information is available to find in a short period. On May 4th, 2022 this hackathon took place: a group of 22 students split up into 11 groups were asked to find the networks of 26 volunteers in as much detail as possible. Each group was given a little less than four hours to find links for seven volunteers that were assigned at random. While some students went to their favourite scraping tools to extract as much public information as possible, others used their sleuthing skills to find relationships between fellow church members or family cats.

In the following weeks, volunteers were given the personal list of found links and were asked to assess them on two criteria: is this link correct (do you know this person), and if so, is the type of link correct, e.g. is this person indeed family. For example, if a colleague was found by the students, respondents could answer yes to both questions if the person found was indeed a colleague, no to both questions if the person was unknown to them, and could answer first yes then no if they knew the person found, but this person was a friend or family member rather than a colleague.

This paper consists of an overview and discussion of the results of this hackathon. Section 2 gives more detail on the assignment given for the hackathon. In section 3 we discuss the obtained data, focussing on which sources and types of links are often used while looking for information online, which sources are seen as more reliable than others, and the effect this has on the networks that are found. In Section 3.5 we will discuss the response by volunteers and the accuracy of the results. In Section 4 we will discuss what this anecdotal evidence might suggest for statistical disclosure control on the publishing of networks.

2 Hackathon assignment

On May 4th, 2022 a group of 22 students from Leiden University, most of them from Computer Science, were asked to use publicly available data to find people in the networks of a group of volunteers who had previously given consent for being researched online. The volunteers were mostly from Statistics Netherlands, who had replied to a post on the internal message board asking for volunteers. Some were found through contacts in the university or personal contacts.

After an explanation of the topic, reason and assignment for the hackathon, students were given the assignment on paper, hints for getting started and rules on which methods were and were not authorized. See Appendix D for the full assignment. They were then asked to pair up and start their search, which was scheduled for four hours. Afterwards, students were supplied with pizza and drinks for their effort.

Willem Alexander		Source: https://www.koninklijkhuis.nl/onderwerpen/geschiedenis/koningen-en-koninginnen/willem-alexander-koning-1967				
Source	Target	Type	Subtype (optional)	Distance	Reliability of link	Source
Beatrix	Willem Alexande	Family	Kind	1	High	https://nl.wikipedia.org/wiki/Koninklijke_familie_van_Nederland
Beatrix	Constantijn	Family	Kind	2	High	https://nl.wikipedia.org/wiki/Koninklijke_familie_van_Nederland
Willem Alexande	Amalia	Family	Kind	1	High	https://nl.wikipedia.org/wiki/Koninklijke_familie_van_Nederland
Amalia	Alexia	Family	Zus	2	High	https://nl.wikipedia.org/wiki/Koninklijke_familie_van_Nederland
Willem Alexande	Maxima	Household	-	1	High	Story
Thom de Graaf	Willem Alexande	Work	Colleague	1	High	raadvanstate.nl
Bert	Ernie	Friends	-	20	High	https://nl.wikipedia.org/wiki/Bert_en_Ernie

FIGURE 1. Example of a filled-in spreadsheet for king Willem Alexander

Each team was given a private Google spreadsheet in which to fulfil the assignment. All students were given an example sheet, with an example filled in as shown in Figure 1, and then seven more empty sheets for their seven randomly assigned volunteers. The form consisted of the following items to fill in per assignment:

- **Source** and **Target**, two separate cells in which the students could supply the two people that were linked. These people could be linked by different types of connections, which could be specified in the next column. Links were considered to be undirected, i.e. the order does not matter and students could swap source and target without consequence. Students were encouraged to both supply direct links, i.e. links where either source or target was the assigned volunteer, and indirect links, i.e. relationships not containing the volunteer, indicated by a higher distance.
- The **Type** of link could then be specified using a drop-down menu, consisting of the eight permissible categories: family, household, work, school, neighbours, co-affiliation, friends or other.
- In the column **Subtype** participants could further explain the relationship between the people in source and target, i.e. give information on what their specific relationship was or through which means they knew each other. Examples of included subtypes are 'member of the same church' or 'old classmates'. The field was an open text field, so students were free to specify the type of relationship in whatever way they saw fit.
- Participants were asked to mention the degree of connectedness by using **Distance**, indicating how many links the target was removed from the person of interest. For example, a family member of a known friend would be given distance 2 if that family member was not connected to the person of interest directly. This cell was open and any distance could be used, although participants would be given a warning if they wrote anything that was not a number between 1 and 5.
- The column **Reliability of link** was used to assess their confidence in the found information. If they were confident in their findings, students were asked to rate the reliability high. If they were unsure or did not trust the source, they were asked to rate the reliability as medium or low. This was a dropdown menu with options high, middle and low.
- Students were also asked to provide the **Source** of the information per link. This could be a general name of the website (i.e. Facebook or Twitter) or a link to a specific page containing the information.

At the end of the hackathon, two teams were chosen for awards. One team had won based on the most found links with a large variety of sources, and the other had won based on 'creativity', by using outside-the-box sources such as Strava.

3 Results

Below we describe in detail the links that were discovered and what network structures these formed, followed by a breakdown by link type and source and an analysis of the reliability of these links.

3.1 Discovered links

In total, more than 5,000 links were found. Links were found for all volunteers, but as shown in Figure 2 the number of links found between the volunteers differed enormously; while for over half of the volunteers less than a hundred links were found, two volunteers had over a thousand links each. Of the 26 volunteers, 21 replied to our request to assess the validity of the inferred links, leaving us with 3,310 assessed links, or about 57% of the total.

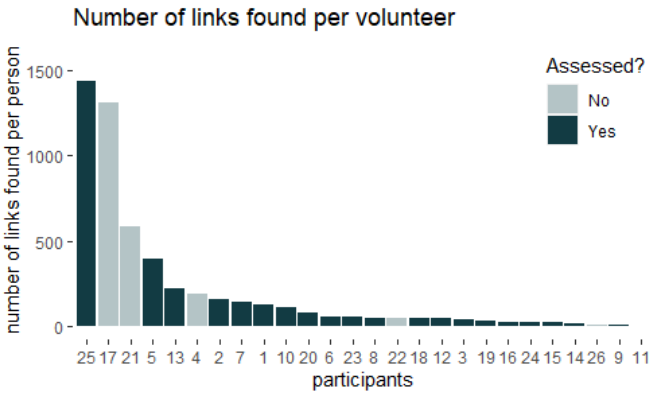


FIGURE 2. The number of found links per volunteer

We noticed during the validation process that an accurate and objective assessment of the networks is complicated for both parties. This created difficulties during the analysis of the data, but in the broader sense, this difficulty extends to potential attackers and could affect the quality of their potential apriori knowledge. Some of the complications we encountered, which might affect the network knowledge of an attacker, were:

- Some links were duplicates, either because multiple teams found the same relationships, or because one individual was seen as multiple people by the students (e.g. A. Nonymous, Annie Nonymous, Dr A. Nonymous, and A.B. Nonymous were seen as 4 people, instead of aliases of one person). Obvious duplicates, such as identical source and target for multiple entries, were easily found and were removed before sending the data to volunteers for validation. When correcting for upper-/lowercase and directionality of the links, roughly 14% of all found links were duplicated between teams. Another 3% of links show up more than once in the network after deduplication due to different link types being assigned. We have not been able to determine the percentage of links that were duplicates due to aliases.
- Some of the links included deceased individuals.
- Interpretation of links and real-life networks can be different between volunteers, students and researchers. Whether a person decides that for example, an old co-worker is still part of their network seems to differ from person to person.
- Some volunteers noticed that while the links found might have been accurate, they were not familiar with the person. This happened frequently with colleagues: while often the work links seemed to have a common publication or LinkedIn match and thus increasing the odds of a participant being familiar with the found link, sometimes unknown or even all publicly available co-workers from the company were listed. This meant participants were unable to accurately assess the validity of the links.
- Students and volunteers had some difficulties with clearly separating the different types of links. Often found links were categorised as friends by the students, while other types such as family, work or school would be more accurate. This is partly because in real life, these networks also overlap: someone can for example be both a friend, a colleague and a household member. Here, both volunteers and students had to choose just one type, thus not allowing for this real-world complexity in relationships. Similarly, the difference between a work and co-affiliation relationship was difficult to discern for the students, causing inconsistent typing of similar links between the teams.

3.2 Link types and sources

There is a clear difference in the number of links found per person, as can be seen in Figures 2. Additionally, there is a large skew in the number of links found for the different types of links, and which sources were consulted during the hackathon. Several sources were used to collect information about the volunteers. Looking at sources that have been used for at least 10 or more links, we can divide them into six categories:

Category	Websites
Social Media	Facebook, Twitter, Instagram, LinkedIn
Personal websites	Medium, personal websites of volunteers
Company websites	Websites of universities (UvA, RUG, CBS, LEI); websites of employers; websites from research projects
Scientific publishing	Researchgate, Google Scholar, GitHub, r-project, frontiersin
Associations, hobbies	Schoolbank, ditismijnteam, lazerlab, creativemornings, BOINC
Other	Miscellaneous websites

As can be seen in Figure 3, there is a clear skew in the types of links and the sources that were used. Most links were found through social media websites and are of the category 'Friends'.

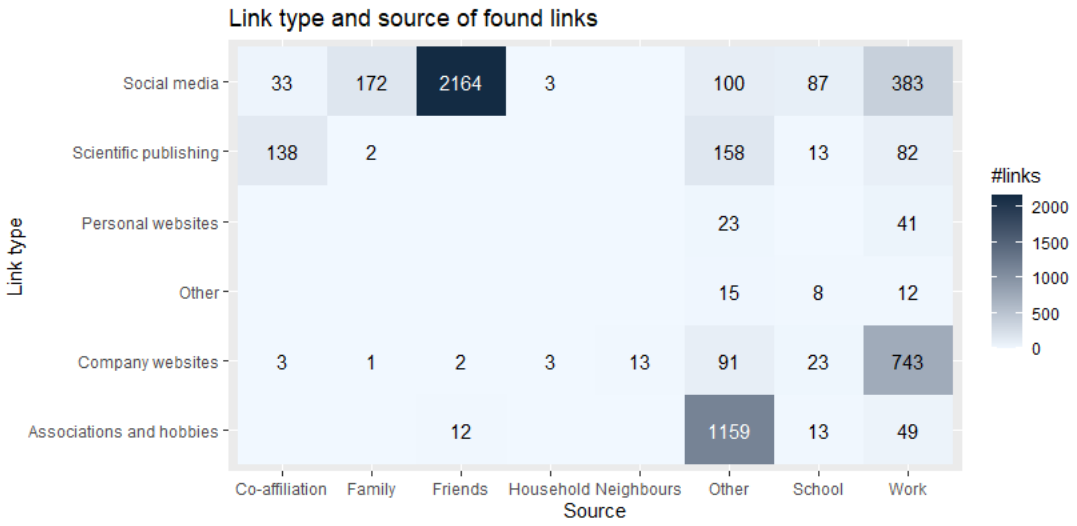


FIGURE 3. Number of links found per type and source

3.3 Discovered network structure

While students were encouraged to find links between neighbours and relationships of neighbours, i.e. higher degree links, most links (90%) were direct links from the target volunteer. The maximum distance assigned to a found link during the hackathon was 3, with 3% of all found links. This suggests students found it easier to find closer links. This could be an unintended effect of our assignment, where given the limited time students preferred to focus only on the first target or where the assignment unintentionally gave more weight to direct links. It is therefore difficult to determine whether this is also a reasonable conclusion for real-world scenarios. As can be seen in Figure 4, which shows the found networks for all Statistics Netherlands (CBS) employees, the found networks indeed mostly resemble hubs with only links of distance one. We find 6 components in the graph: there is one major component, consisting of 1,268 links, and five smaller components, from size 55 to size 2. One might assume the components are caused by the fact that all volunteers work for the same company, but not all of the interconnectedness can be explained by links of type 'work' or 'co-affiliation'; if we only look at these two types of links we are left with 10 components, six of which are part of the largest component in the full network.
It is difficult to discern whether the networks are truly connected, both in the real world and the data set: some components are caused by volunteers finding links with the same name, for example in the family networks, and it is difficult to discern whether these names refer to the same person. For a deeper dive into the networks found per type of link, see Figure 11 in Appendix A.

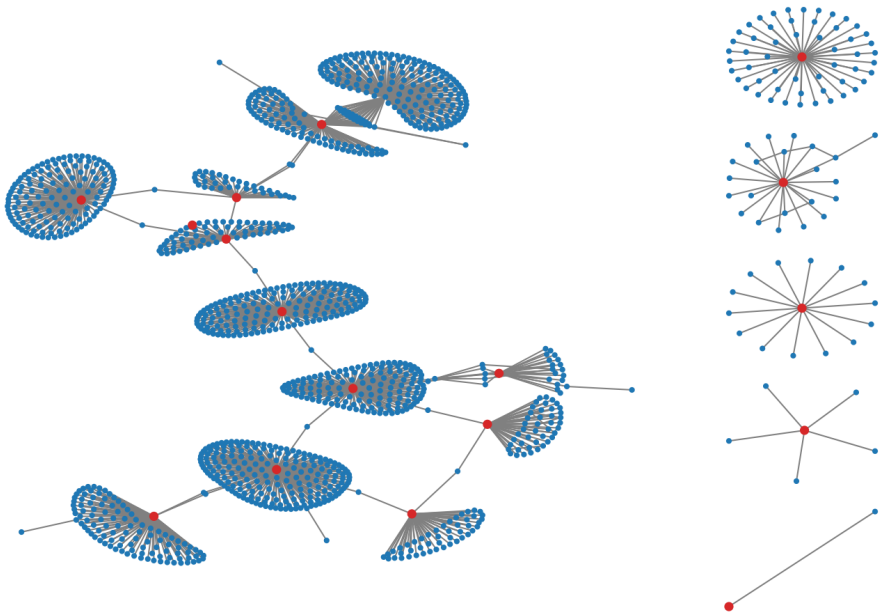


FIGURE 4. Networks of CBS employees, red dots represent volunteers

3.4 Perceived reliability of links

Students were also asked to rate their confidence in the reliability of their found links. We can see clear differences between the different types of sources: where information found on personal and company websites is deemed credible, the information found through social media is decidedly less confidence-inspiring. Interestingly, information found through websites about associations or hobbies, like websites collecting information about sports teams or previous classmates from elementary school, was universally regarded as questionable sources. See Figure 5 for the students' perceived reliability of the sources used during the hackathon.

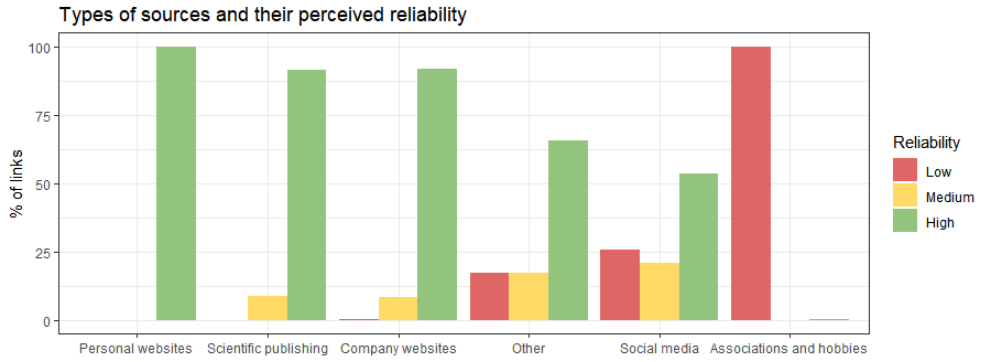


FIGURE 5. Students' perceived reliability of the types of sources used for the hackathon, all links

When looking at the perceived reliability of information on the internet, it seems sensible to look to Figure 5 for a general conclusion. This figure also seems to match intuition: it seems reasonable that one looking to expose the personal relationships of a volunteer is more likely to trust personal websites that the volunteer himself provides information on, rather than websites such as Facebook or indeed possibly outdated information about club memberships. However, seeing as due to non-response from some of the volunteers we were unable to

perform the analyses on all data, we also looked at the perceived reliability of only the assessed links in the data. Figure 6 is a good reference for the data that will be analysed in the following sections.

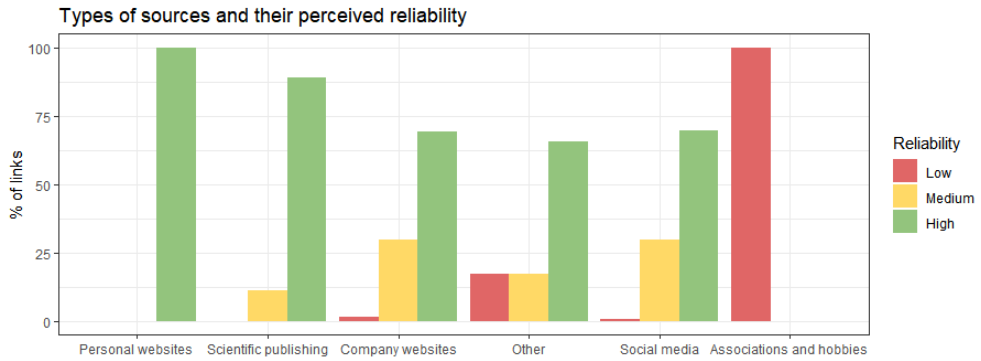
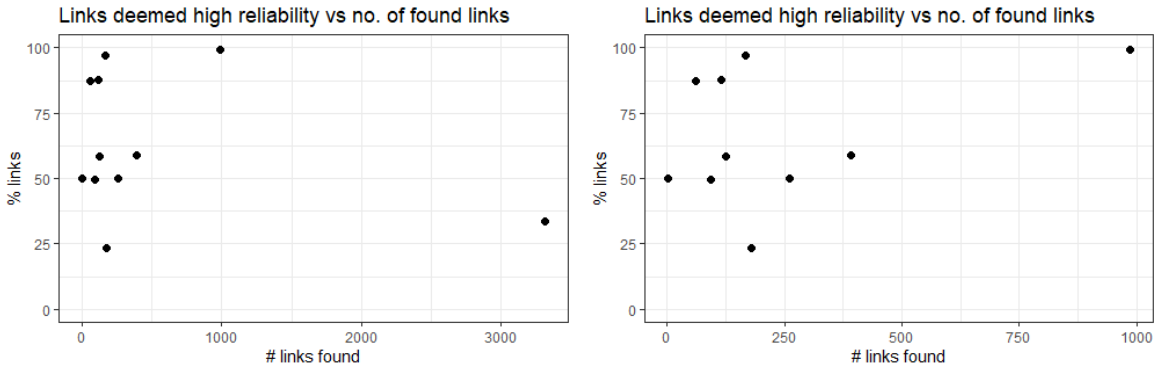


FIGURE 6. Perceived reliability of the types of sources used for the hackathon, assessed links

Figure 6 clearly shows the impact of not including the unassessed links: we see here that contrary to the full data set, social media is deemed to be much more reliable by the students, increasing from 53% high confidence to almost 75%. This is a big shift but can be explained by the size of the unassessed networks: of the five volunteers that did not get back to us about their results, two of them were in the top three largest networks found. All unassessed links in total counted for almost half of all links (45%) found through social media, and more than two-thirds (72%) of all links found through company websites. While it is obvious the exclusion of these links has the potential to alter the outcome significantly, it is not necessarily obvious why they caused this shift.

One might assume that the number of found nodes per team has an impact on the perceived reliability: when there is less time per link to assess the validity, it might lead to less confidence in the found information.



(A) Number of found links vs. confidence, all teams

(B) Excluding team with >3000 links

FIGURE 7. The number of found links per team and the percentage of links deemed reliable.

However, when we look at the number of found links per team and the percentage of links deemed reliable in Figure 7, we see no clear correlation between these two data points. Figure 7a show us no clear-cut correlation between 0 and 500 found links, a maximum percentage of links deemed highly reliable at 1000 found links, and back down to an almost minimum of around 30% at 3300 found links. Even when excluding the team with more than 3000 links from the plot, Figure 7b does not show a clear pattern. A definite conclusion can therefore not be drawn from this data; it is more likely that the method of finding the links, e.g. whether scraping tools are used and the confidence in those tools, is a bigger factor in deciding the reliability of the link.

3.5 Response from volunteers

Volunteers were asked two questions when assessing the validity of the links. First: is this link correct, meaning is this a person you know and is part of your network? Second: is the type of link correct, meaning for example is the person indeed a family member if it is noted as a family member?
When analysing the response, we see that 1,670 (50.5%) of the assessed links are completely correct, meaning the volunteer knew the person and the type was correctly identified (Yes/Yes). 1,356 links (41%) were completely incorrect, meaning the person was unknown to the volunteer and thus type was also incorrect (No/No). For another 241 links (7.3%) the volunteers knew the person, but the type of link was wrong (Yes/No). This happened mostly for links found for the category 'Friends': often these were scraped from Facebook and therefore labelled as friends, but in many cases, volunteers found labels like 'Family' or 'School' to be more appropriate. Curiously, we also found 42 links for which the volunteers noted that while the link was unknown to them, the type of link was correct. For most links, this happened to be from one volunteer for which students from the same course and starting year were found. The volunteer knew some of the links found, but not all of them, and therefore chose No/Yes for all links of that type.
The consensus between the volunteers is that the results are different than they expected. Often volunteers expressed surprise at the number and types of links found; either fewer links were found than predicted, or different, seemingly less important links were found. For some, students were able to produce quite accurate and complete sub-networks, e.g. for family or colleagues, but often the networks resembled an arbitrary selection of relationships, where for example the networks only contained a father but no mother or siblings, despite them being part of their network and visible online. See Appendix B for a summary of the responses received.

3.5.1 Accuracy of the links. For most link types, there were more correct inferences than incorrect ones. However, for the category household and neighbours, we find that most links are completely incorrect. Because the number of links found in this category is significantly smaller than in all other categories, it seems quite difficult to find information about these relationships online. It must be noted that most links in these categories were incorrect due to a case of mistaken identity, and thus we are unsure of the validity of the information.

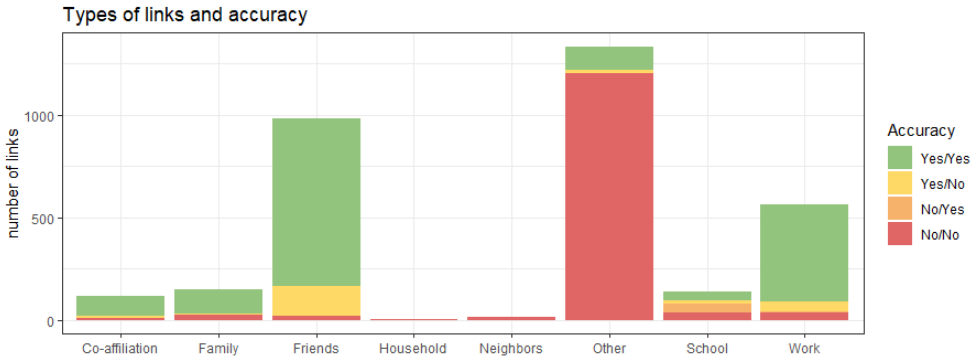


FIGURE 8. Accuracy of the inferred links per category

When looking at Figure 8, there are some details of note. Firstly, almost all the links found in the category 'Other' seem to be incorrect. This is mostly due to the links found for one volunteer: about 78 % of all links in this category were scraped from one online service website that a volunteer shared with a community, of which the volunteer knows no one. If we remove these links, the percentage of incorrect links goes down to about 50%. While a significant drop compared to the accuracy in Figure 8, this is still a relatively high number of incorrect links compared to other categories. Interestingly, 'School' is another category that does not seem to be performing well.
When looking at the type of source used to derive a link, we can see that certain sources produce more accurate links than others. See Figure 9 for all the categories. Especially personal websites, while only accounting for 38 links, seemed to be the most reliable: all of the links found through these websites were correct. In stark

contrast, we see that hobby websites produce highly unreliable results. Curiously, while social media was the third worst in terms of perceived reliability, they did contribute to many correct links.

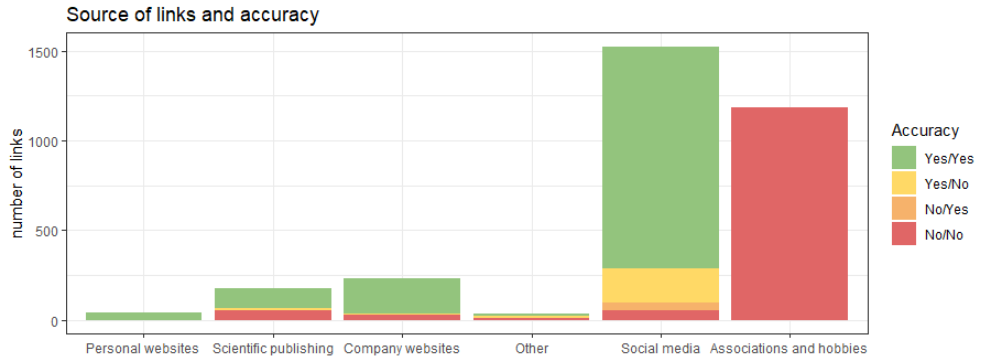


FIGURE 9. Accuracy of the inferred links per source

3.5.2 Accuracy versus perceived reliability. It seems that students are mostly able to assess the accuracy of their information quite well. If we look at the perceived reliability of the information and the actual accuracy of the found links, we see that most incorrect data was indeed labelled as unreliable, while most of the correctly inferred links were deemed reliable. See Figure 10: when students were unsure about their inference, it turns out most of them were indeed completely incorrect. For medium and high perceived reliability the numbers show more variation, but in most cases the links and type were correct.

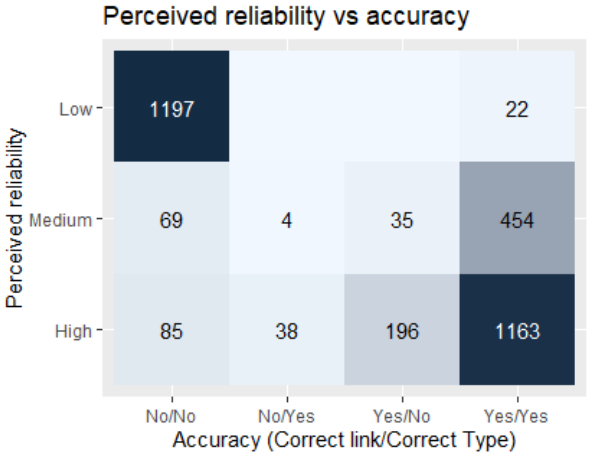


FIGURE 10. Accuracy of the links compared to perceived reliability by students

It is interesting to see if there are differences to be found in the categories and the types of sources used. Figure 12 in Appendix A shows us which sources enjoyed a false sense of reliability, and which sources were more likely to be incorrectly perceived as unreliable. Most sources have a majority of links that have high perceived reliability and are correct in both link and type, i.e. were correctly assumed to be credible. Personal websites are the most trustworthy source, with all links correct and given high reliability. For associations and hobbies, the converse is true; all links for these sources are correctly deemed questionable.

		Social Media			Scientific Publishing		
		Perceived reliability			Perceived reliability		
		High	Medium	Low	High	Medium	Low
Accuracy	Yes/Yes	827	375	8	96	14	0
	Yes/No	159	31	0	7	4	0
	No/Yes	38	2	0	0	0	0
	No/No	16	36	1	51	1	0

TABLE 1. Number of found links and perceived reliability by students versus accuracy (correct link/correct type), for sources Social media and Scientific publishing

Table 1 shows the numbers for two sources: social media and scientific publishing. Social media has a pretty common spread: most links are correct and given high reliability, and even many links with medium reliability

are correct. Some of the links are incorrect and falsely given high or medium reliability, but this is a small portion of all links. An interesting category is 'Scientific Publishing', where a relatively high percentage of links deemed highly reliable were incorrect.
When looking at the difference in link type, as can be seen in Figure 13, again we find most links being given high reliability and assessed to be fully correct. Friends is a good example of this: as can be seen in Table 2, 61% of the links in this category are indeed of this type.

		Friends			School		
		Perceived reliability			Perceived reliability		
		High	Medium	Low	High	Medium	Low
Accuracy	Yes/Yes	599	220	0	34	5	1
	Yes/No	128	14	0	15	2	0
	No/Yes	0	1	0	38	0	0
	No/No	2	6	12	13	2	23

TABLE 2. Number of found links and perceived reliability by students versus accuracy (correct link/correct type), for friends and school links

In contrast, links found for school relationships are less reliable; even when students were confident in their inference and gave them high reliability, links are more likely to be partially or completely wrong than completely correct. These numbers show us that in most cases, students are competent in assessing the validity of the links; when students rated their links as low reliability, they were almost always incorrect, and when they were rated medium or high reliability, they were often correct. The only exceptions, besides school, are links of type neighbour or household: even though these were given medium or high reliability, they were almost always incorrect.

4 Discussion, conclusion and outlook

These findings, while a small test case, yield some interesting results that can lead us to some tentative conclusions. Before we do this, we would like to highlight some of the difficulties of generalizing these results. It is important to note that while students had four hours to work on the assignment, they were each given seven volunteers to find information on. Students were encouraged to find as many nodes in their networks as possible, thus possibly focusing on people that were easiest to find online. It is therefore unlikely that the found networks are a good reflection of all that is available online for each volunteer, especially at higher distances.
Furthermore, there was a significant variation in the methods employed by students. Some conducted detailed manual searches of volunteers' online social profiles, resembling an attacker gathering apriori information about a specific individual. Others opted for a broader approach, scraping from platforms like Facebook and LinkedIn, similar to attackers searching for vulnerabilities in a data set.
Finally, neither the volunteers nor the students were a good representation of the general public. A lot of the volunteers were researchers that were easily found through research-related websites. Furthermore, many of the students were from the faculty of computer science, which means we expect them to have a specific skill set that is not necessarily presumed for the general public.

4.1 Conclusions

When looking at this data set, there are some interesting patterns to find in which sources and types of links are easily found and assessed correctly. Information has been found for all volunteers, albeit for some decidedly more than others. We are most interested in the types of links: which layers of the population network are easier to reveal, and might therefore need better protection?

4.1.1 Which information is easy to obtain? The two categories for which most nodes were found, excluding the 'Other' category, are friends and colleagues. This is likely due to several factors:

1. Figure 3 demonstrates that social media is the most used source: of the four social media websites, most links were found on LinkedIn and Facebook. LinkedIn primarily offers work-related connections, while Facebook encompasses a broader range where users connect with friends, family, and acquaintances. The prevalence of social media as a source could be attributed to its accessibility and ease of scraping, or it could be due to its widespread usage among the general population, making it a natural starting point for research. It is difficult to determine whether the abundance of links found through these websites is a result of their accessibility or if the students simply prioritized searching on these platforms.
2. The category 'Friends' in particular is quite broad and can encompass a wide range of interpersonal links. Because of its vagueness, links that might have been better suited for categories such as 'School' or 'Family' were often categorised as 'Friends' if participants were aware of the link, but unsure of the true nature of the relationship.

Nevertheless, it seems the found networks for these types are mostly correct, thus suggesting especially these kinds of relationships are vulnerable to disclosure.

4.1.2 Which information is difficult to obtain? When looking at the seven used categories, barely any information was found on household and neighbour-type links. Coupled with the low accuracy of the assessed links, it seems clear that this information is in general hard to find online. While information about (old) school relationships seems a bit easier to find, as there are a little over 100 assessed links of this type, very few of them are correct. This is even the case when students think their information is reliable: only 34 of the 100 links with high reliability are completely correct.
Most students looked only at first-order relationships, and thus the found networks mostly resembled hubs with isolated nodes. This seems to suggest that finding higher-order relationships is more difficult, but this focus on first-order relationships could also be an unintended artefact of our assignment.

4.1.3 How accurate are the assessments? Most students show a good judgement of the credibility of online information, as evidenced by Figure 10. The majority of links are either assigned low reliability and proven incorrect, or assigned high reliability and validated as correct. This pattern applies to most source types and link categories, with the exceptions being 'School', 'Household', and 'Neighbours'.

4.2 Questions left for further research

The main aim of this research was to get a better insight into the ease with which certain personal information can be found on the internet. This insight will help us get a better understanding of what information is more vulnerable to disclosure when combining data from national statistical agencies with outside sources. This hackathon is just a starting point for this topic; more research is needed to get a fuller picture of the possible effect of online research.
This leads to another open question, namely how to approach these possible vulnerabilities from a statistical disclosure control standpoint. Outside sources are not always taken into account in traditional statistical disclosure control. Knowing what information is vulnerable is therefore a good first step to understanding the issue. In general, the question of how to involve public information in statistical disclosure control is still open for discussion. Third, the initial findings on what link types are typically easy to discover through social media and which ones are not, might be a starting point for more automated approaches towards understanding how accurately population-scale social network ties derived from register data reflect actual social ties.
This hackathon specifically looked at networks of Dutch citizens. We intend to study which adaptions can be made to our current risk assessment for statistical disclosure control regarding publishing networks. Furthermore, this research ties into previous research de Jong et al. (2023a); van der Loo et al. (2021); van der Loo (2022) to look into the vulnerabilities of the network structure itself. This research is still ongoing and we aim to include the results of this hackathon in further research, for example which anonymisation methods are most effective.

References

Aliprandi, C., J. Irujo, M. Cuadros, S. Maier, F. Melero, and M. Raffaelli (2014, 06). Caper: Collaborative information, acquisition, processing, exploitation and reporting for the prevention of organised crime. Volume 434.

Bokányi, E., E. M. Heemskerk, and F. W. Takes (2023). The anatomy of a population-scale social network. *Scientific Reports 13*(1).

Centraal Bureau voor de Statistiek (2022, Oct). Internettoegang en internetactiviteiten; persoonskenmerken. https://www.cbs.nl/nl-nl/cijfers/detail/84888NED.

Confessore, N. (2018, Apr). Cambridge analytica and facebook: The scandal and the fallout so far. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.

de Jong, R. G., M. P. J. van der Loo, and F. W. Takes (2023a, jun). Algorithms for efficiently computing structural anonymity in complex networks. *ACM J. Exp. Algorithmics*.

de Jong, R. G., M. P. J. van der Loo, and F. W. Takes (2023b). Beyond the ego network: the effect of distant connections on node anonymity. *arXiv preprint 2306.13508*.

de Vries, M., P.-P. de Wolf, and M. van der Loo (2021). Statistische beveiliging, online privacy en osint. Unpublished internal report.

Khanna, P., P. Zavarsky, and D. Lindskog (2016). Experimental analysis of tools used for doxing and proposed new transforms to help organizations protect against doxing attacks. *Procedia Computer Science 94*, 459–464. The 11th International Conference on Future Networks and Communications (FNC 2016) / The 13th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2016) / Affiliated Workshops.

Koot, G., M. Huis in 't Veld, J. Hendricksen, R. Kaptein, A. Vries, and E. van den Broek (2014, September). Foraging online social networks. In M. den Hengst, M. Israël, D. Zeng, C. Veenman, and A. Wang (Eds.), *Proceedings of the 2014 IEEE Joint Intelligence and Security Informatics Conference (JISIC2014)*, pp. 312–315. IEEE. 10.1109/JISIC.2014.62 ; null ; Conference date: 24-09-2014 Through 26-09-2014.

Pastor-Galindo, J., P. Nespoli, F. Gomez Marmol, and G. Martinez Perez (2020). The not yet exploited goldmine of osint: Opportunities, open challenges and future trends. *IEEE Access 8*, 10282â10304.

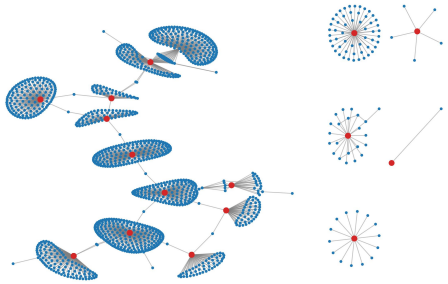
van der Laan, J., E. de Jonge, M. Das, S. Te Riele, and T. Emery (2022). A whole population network and its application for the social sciences. *European Sociological Review 39*(1), 145â160.

van der Loo, M. (2022). Topological anonymity in networks. Technical report.

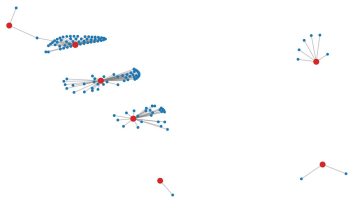
van der Loo, M., R. de Jong, F. Takes, M. de Vries, and P.-P. de Wolf (2021). Structural uniqueness in networks. Expert Meeting on Statistical Data Confidentiality.

Appendix A Figures

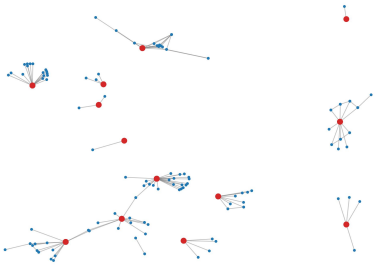
(A) All found links for CBS employees



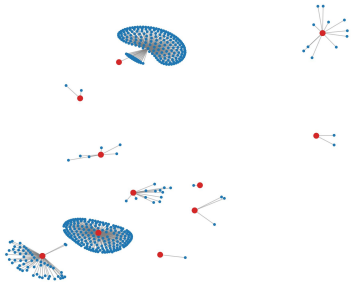
(B) Subgraph of inks with type 'Co-Affiliation'



(C) Subgraph of inks with type 'Family'

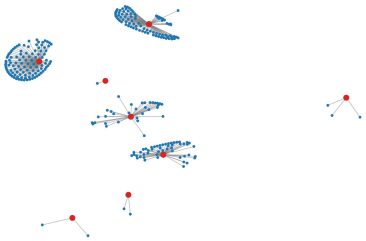


(D) Subgraph of inks with type 'Friends'

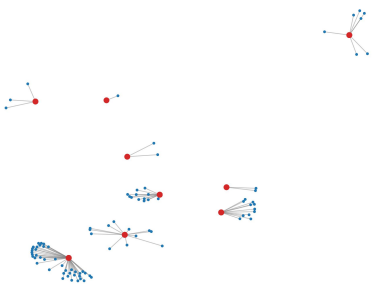


(E) Subgraph of inks with type 'Household'

(F) Subgraph of inks with type 'Other'



(G) Subgraph of inks with type 'School'



(H) Subgraph of inks with type 'Work'

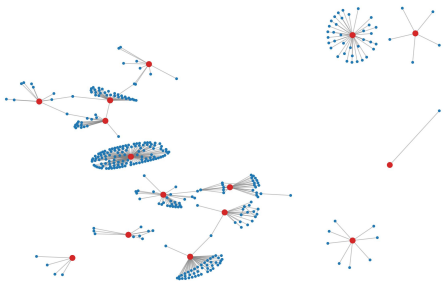


FIGURE 11. Networks of CBS employees, red dots represent volunteers. No links of type 'Neighbour' were found for this subset of volunteers.

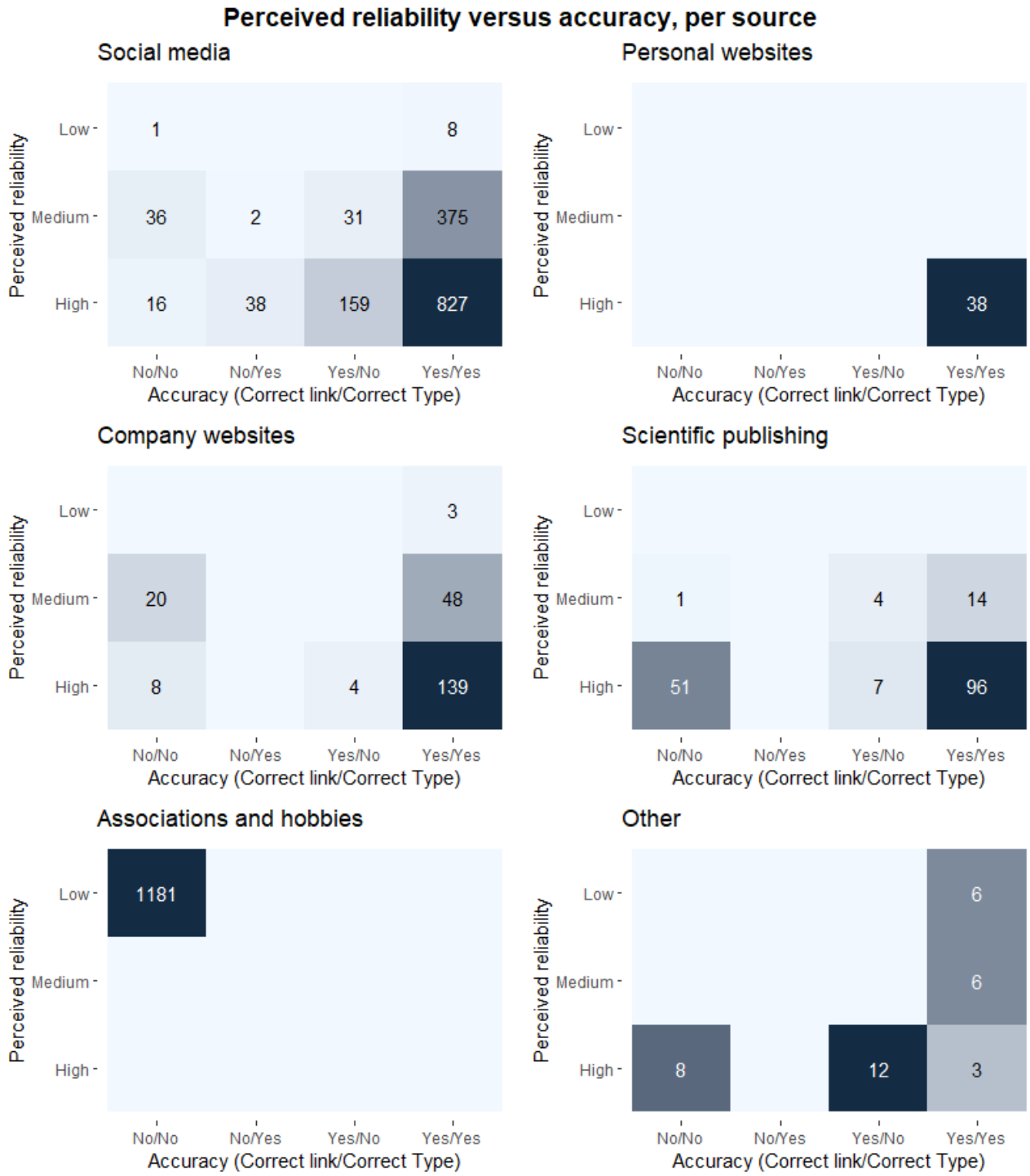


FIGURE 12. Accuracy of the links compared to perceived reliability by students, by source

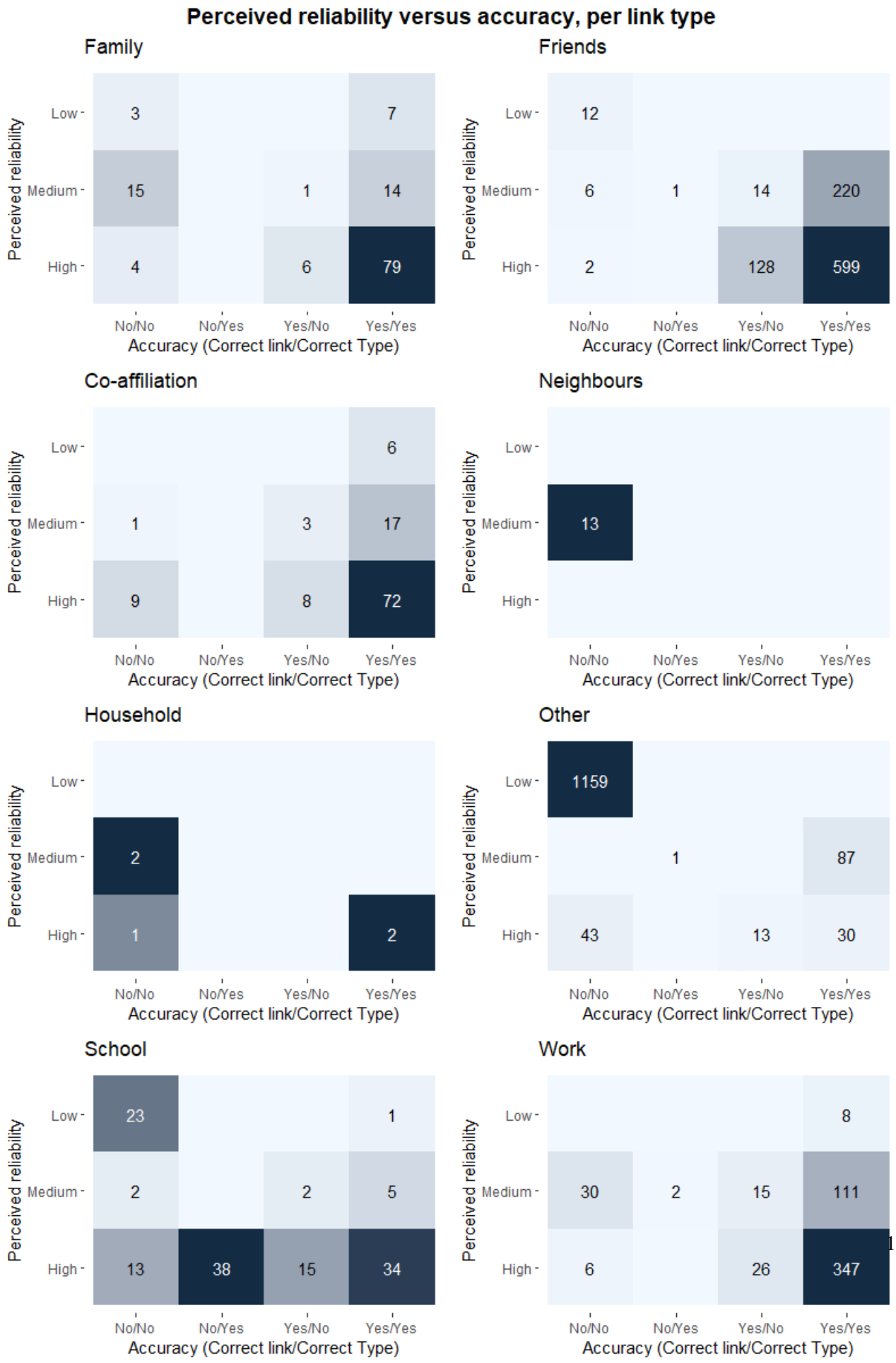


FIGURE 13. Accuracy of the links compared to perceived reliability by students, by link type

Appendix B Responses from volunteers

Participant	No. of links	Most found	Comment
2	153	Other	"I'm impressed by the quantity but not the quality of the links. I don't think researchgate is a good source, because it mostly consists of cited scientists or people working for my company or the university, and most of these I really do not know. Otherwise, a couple of 'friends' have been found that I'm not friends with, while some of my best friends haven't been found at all. I expected them to be able to find my network in more detail, but if I were to do it myself, I would find different links."
3	38	Work	"This is a remarkable selection of people from my networks, which make me wonder why those networks didn't reveal more people. For example, why is only my dad revealed through [genealogy website] and no other family? A lot more people are revealed there. And the friends found on Facebook is a very minimal selection. I cannot place exactly why these people from work would be found and others wouldn't be. Most of the found links I know very superficially, while colleagues I talk to daily are not included. Also, there should be more sources available, such as [schoolbank] for school related links. Other sources have outdated information available, but they do reveal new sources or can help with fact checking the links."
5	399	Friends	"Many links from Facebook, but that makes sense. I noticed not many links were found through LinkedIn, despite me having quite some connections there. Most of the people found as friends are Facebook friends, but I wouldn't call them friends in real life, as most of them are acquaintances, neighbours, or (ex) family in law."
6	55	Work	"Some of the links found through LinkedIn are friends, not colleagues."
9	5	Work	"I expected this to come up, this is one of the few publications from my previous job. My network and family is quite big, so I'm surprised with the lack of links found. If I google myself, I am able to find more information."
10	306	Friends	"All nodes for work are related to an old publication, so perhaps the label 'Work' is no longer applicable to those that have left the company."
11	1	Work	"They found even less than I expected. This colleague hasn't been a direct co-worker of mine for years. We still work for the same company, but I have tens, if not a hundred closer colleagues."
16	29	Family	"Not everything was correct, but most of it was. Apparently my family is easy to track online"
18	47	School	"Surprising to see how easily you could find links and how accurate you are"
25	1460	Other	"Most of the links are from an online service which I share with a community, of which I know no one. All Facebook links noted as 'Friends' were technically correct, but Family/School/etc would often be a better fit."

Appendix C Tables

		Work			Friends			Family		
		Perceived reliability			Perceived reliability			Perceived reliability		
		High	Medium	Low	High	Medium	Low	High	Medium	Low
Accuracy	Yes/Yes	347	111	8	599	220	0	79	14	7
	Yes/No	26	15	0	128	14	0	6	1	0
	No/Yes	0	2	0	0	1	0	0	0	0
	No/No	6	30	0	2	6	12	4	15	3

		Other			Co-affiliation			Household		
		Perceived reliability			Perceived reliability			Perceived reliability		
		High	Medium	Low	High	Medium	Low	High	Medium	Low
Accuracy	Yes/Yes	87	0	13	72	17	6	2	0	0
	Yes/No	13	0	0	8	3	0	0	0	0
	No/Yes	0	1	0	0	0	0	0	0	0
	No/No	43	0	1159	9	1	0	1	2	0

		School			Neighbours		
		Perceived reliability			Perceived reliability		
		High	Medium	Low	High	Medium	Low
Accuracy	Yes/Yes	34	5	1	0	0	0
	Yes/No	15	2	0	0	0	0
	No/Yes	38	0	0	0	0	0
	No/No	13	2	23	0	13	0

TABLE 3. Number of found links per category and perceived reliability by students, versus accuracy

		Social Media			Personal websites			Company websites		
		Perceived reliability			Perceived reliability			Perceived reliability		
		High	Medium	Low	High	Medium	Low	High	Medium	Low
Accuracy	Yes/Yes	827	375	8	38	0	0	139	48	3
	Yes/No	159	31	0	0	0	0	4	0	0
	No/Yes	38	2	0	0	0	0	0	0	0
	No/No	16	36	1	0	0	0	8	20	0

		Scientific publishing			Associations and hobbies			Other		
		Perceived reliability			Perceived reliability			Perceived reliability		
		High	Medium	Low	High	Medium	Low	High	Medium	Low
Accuracy	Yes/Yes	96	14	0	0	0	0	3	6	6
	Yes/No	7	4	0	0	0	0	12	0	0
	No/Yes	0	0	0	0	0	0	0	0	0
	No/No	51	1	0	0	0	1181	8	0	0

TABLE 4. Number of found links per source and perceived reliability by students, versus accuracy

Appendix D Hackathon assignment

Centraal Bureau voor de Statistiek

POPNET

Universiteit Leiden The Netherlands

University of Amsterdam

HACKATHON: FROM PERSON TO OPEN DATA

3 May 2022

ORGANIZED BY:

ANO-NET

STATISTICS NETHERLANDS (CBS)

POPNET

CNS GROUP OF LEIDEN UNIVERSITY

From person to open data:
How anonymous are you?

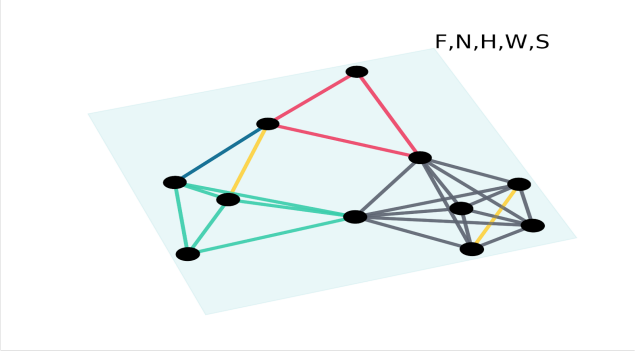
Almost everyone is part of several social networks, each network being formed by its own type of link. For several years now, Statistics Netherlands (CBS) has also been conducting research into the Netherlands modeled as a network, including with the POPNET project. In the context of this project, research is being done into properties of, for example, the Dutch family network, the colleague network, and the neighbor network.

On the one hand, these networks are very interesting for researchers, on the other hand, CBS is prohibited by law from publishing data that can be traced back to individuals (or companies). In this context, it is interesting to know to what extent network data can be derived from public data, because that says something about the extent to which people in networks are anonymous.

Assignment

In this assignment you will try to find out the networks of a number of volunteers based on publicly available data. Link types could represent anything and any type of link may be included. The types that we are especially interested in are:

- Family relations (parent-child, or other)
- Household
- Work relations (same employer)
- Classmates or same school
- Neighbors or living in the same neighborhood
- Other co-affiliations
- Friends



Additionally, finding the network of an individual means not just the direct connections to a neighbor, but also "links between neighbors" or "neighbors of neighbors", "neighbors of neighbors of neighbors", and so on. Also, note that it is also possible for multiple link types to exist between the same two persons.

Fulfillment

Each team receives a unique link to a google sheet (see end of this document) where an example is included with some comments. Each tab of this sheet will contain the network of the given person. These sheets contain the following columns:

- **Source, target:** the link found (you can assume that all links are undirected)
- **Type:** is in one of the given categories (family, household, ...) no other options can be used.
- **Subtype:** any additional information about the linktype, such as parent-child for family. This field is optional.
- **Distance:** the (estimated) distance from the furthest node to the volunteer i.e. if the volunteer occurs in the link, the distance equals one edge. Note that the distance should be no larger than 5, otherwise a small error is given.
- **Reliability of link:** your estimation of the reliability of the link, how certain are you that this link is real?
- **Source:** the source(s) used to derive this link. When a specific link can be derived from multiple sources, please mention all sources, but note that this counts as one link.

Additionally, we ask every team to keep track of a small logbook (.txt format) where you keep track of which types of sources are used. This should contain:

- A summary of which (distinct) sources are used (+ the number of sources)
- Which other tools are used (if you have written any code, please mention this and include the code in your submission as well)

Assessment

The assessment takes into account:

- The number of direct links to volunteer (extra points for specific linktypes)
- The number of indirect links (distance>1 from volunteer)
- Different types of links found (see above)
- The number of different sources used (mentioned in the logbook)

Method

You may only use publicly available information via the internet. Techniques that amount to computer intrusion[1] (hacking), information buying, phishing or other forms of social engineering are not allowed. The latter also includes establishing connections on social media with the aim of revealing links. Teams using an unauthorized technique will be disqualified.

[1] https://www.om.nl/onderwerpen/cybercrime/hack_right/wetsartikel-computervredebreuk

Results

The results are for research purposes only and may not be disseminated further. Access to the spreadsheets used will be restricted after the hackathon. The (alleged) information you collect about people will be treated as confidential and may not be used or spread outside the hackathon.

Hints

- **Search engines** such as: Google, bing, DuckDuckGo, Yahoo ...
- **Socials**: LinkedIn, facebook, instagram, twitter, reddit, youtube, ...
    - If you know a user name, you can find on which platforms this person can be found with: https://instantusername.com/#/
    - Archive.org / Hyves.nl
- **Services:** github, gitlab, stackoverflow, marktplaats, vinted, skillshare ...
    - Forums: fok, girlscene,
- **Business related:**
    - Kamer van Koophandel / OpenKvK
    - Opencorporates
- **Other**:
    - https://osintframework.com/ gives access to many different tools and websites. (Note that the focus of this tool is on the American population)

Link to spreadsheet: https://tinyurl.com/sxv2fzpp *(Team 1)*
Please send your logbook to: popnet@uva.nl