Running head: ONLINE SOCIAL NETWORK DATA AS SOCIOMETRIC MARKERS

Online Social Network Data as Sociometric Markers

Jens F. Binder, Sarah L. Buglass, Lucy R. Betts, and Jean D.M. Underwood

Nottingham Trent University

PRE-PUBLICATION VERSION

IN PRINT WITH AMERICAN PSYCHOLOGIST (http://www.apa.org/pubs/journals/amp/)

Author Note

Jens F. Binder, Sarah L. Buglass, Lucy R. Betts, Jean D.M. Underwood, Division of Psychology, School of Social Sciences, Nottingham Trent University, UK.

Correspondence concerning this article should be addressed to Jens Binder, Division of Psychology, Nottingham Trent University, Burton Street, Nottingham, NG1 4BU, UK. Email: jens.binder@ntu.ac.uk. Phone: +44 115 8482416.

Abstract

Data from online social networks carry enormous potential for psychological research, yet their use and the ethical implications thereof are currently hotly debated. The present work aims to outline in detail the unique information richness of this data type and, in doing so, to support researchers when deciding on ethically appropriate ways of collecting, storing, publishing, and sharing data from online sources. Focusing on the very nature of social networks, their structural characteristics and depth of information, a detailed and accessible account of the challenges associated with data management and data storage is provided. In particular, the general non-anonymity of network data sets is discussed, and an approach is developed to quantify the level of uniqueness that a particular online network bestows upon the individual maintaining it. Using graph enumeration techniques, it can be shown that comparatively sparse information on a network is suitable as a sociometric marker that allows for the identification of an individual from the global population of online users. The impossibility of anonymizing specific types of network data carries implications for ethical guidelines and research practice. At the same time, network uniqueness opens up opportunities for novel research in psychology.

Keywords: social network analysis, social network sites, anonymization, research ethics, graph enumeration

Online Social Network Data as Sociometric Markers

Ever since the dramatic rise of online social network sites (SNS; boyd & Ellison, 2008) and their subsequent impact on research (Wilson, Gosling, & Graham, 2012), debates have sprung up regarding ethically appropriate ways of obtaining and managing digitally derived network data for research purposes. An understanding of what is and is not appropriate in handling specific types of online data is vital if psychologists and researchers from neighbouring disciplines want to utilise this unprecedented wealth of information.

It is generally accepted that research opportunities in the online domain bring about novel challenges in terms of research ethics, and this is reflected in ongoing efforts to amend existing ethical codes of conduct (British Psychological Society, 2013; Fiske & Hauser, 2014; Kraut, Olson, Banaji, Bruckman, Cohen, & Couper, 2004; National Research Council, 2014). While the most prominent debates and arguments have centred on the process of data collection, e.g., issues of obtaining informed consent from potentially unsuspecting users of SNS (Gleibs, 2014; Fiske & Hauser, 2014; Moreno, Goniu, Moreno, Diekema, 2013; Puschmann & Bozgad, 2014; Shah, Cappella, & Neuman, 2014) in field studies with or without experimental manipulation, fewer arguments have focused on the actual properties of network data. Specifically, the highly characteristic structure mapped out in a social network can render anonymization efforts ineffective (Hay, Miklau, Jensen, Towsley, & Weis, 2008; Narayanan & Shmatikov, 2009; Zimmer, 2010) and as such privacy protection for the people forming the contacts within the network can no longer be guaranteed.

This paper provides an argument for a more general non-anonymity in SNS data and to raise awareness of the magnitude of the resulting challenges and opportunities among researchers. To this end, we will demonstrate the considerable information richness that can easily be obtained from users of SNS services. We do so by using graph enumeration

techniques that are not normally considered in psychology and other behavioural sciences. Our analyses highlight, in a way that is hopefully accessible to a wider range of academic disciplines and researchers, in what ways SNS data are truly complex, why this should inform ethical practice in research, and where the wider implications of the use of digital network data lie. The extreme uniqueness of SNS data, according to our argument, forms a highly diagnostic sociometric marker of the network holder, even in the absence of most other contextual information. This marker is akin to a digital signature or "digital DNA" and necessitates additional consideration on the side of the researcher when storing, publishing, or sharing data regardless of attempts at anonymization.

In the following, we will describe the data structure of online social networks and the role this type of data plays in current research. We will then turn to the general uniqueness problem. The core part of the present work is a computational argument and simulation showing that a comparatively simple network structure, as used in recent studies, carries information sufficient to place an individualised sociometric marker on most users in the global population. This approach provides some simple means of assessing the general level of data uniqueness and relates this level to practical recommendations.

Current perspectives on SNS research

The research history of SNS spans barely more than a decade (boyd & Ellison, 2008; Wilson et al., 2012) and is dynamically expanding. The potential that SNS data hold for research has been realised from early on (Lewis, Kaufman, Gonzalez, Wimmer, & Christakis, 2008), and network statistics have been presented that amply demonstrate the impressive amount of information amassed on these sites (Ugander, Karrer, Backstrom, & Marlow, 2011). Online network processes and structures have become increasingly important concepts to a wide range of research areas, from word-of-mouth marketing (Brown, Broderick, & Lee, 2007) to

collaborative learning (de Laat, Lally, Lipponen, & Simons, 2007) to online vulnerability (Buglass, Binder, Betts, & Underwood, 2016) and psychological well-being (Brooks, Hogan, Ellison, Lampe & Vitak, 2015). SNS are of particular interest in the online domain since they contain digital information to map out both non-centred, or whole, networks and ego-centric networks, or ego-nets. The main focus here is on such ego-nets, composed of one ego and alters as nodes, the alter-alter ties and, optionally, ego and alter attributes (Crossley, Bellotti, Edwards, Everett, Koskinen, & Tranmer, 2015). Psychologists are typically interested in such personal networks as attributes of individuals, in studies that relate network characteristics to outcome variables at the individual level (e.g., Brooks et al., 2015; Buglass et al., 2016).

Obtaining such data is now within easy reach of social and behavioural scientists not traditionally concerned with digital data extraction techniques.[1] However, the research history to date has been fraught with difficulties and debates on ethical principles (see, for example, Fiske & Hauser, 2014; Gleibs, 2014; Moreno et al., 2013; Zimmer, 2010). As studies switch from a physical to a virtual environment, parts of the codes of research ethics that have been developed on the basis of direct or human-mediated contact between researcher and participant require modification. While the practical importance of research ethics varies from discipline to discipline, it lies at the heart of psychology as the science of human behaviour, and core debates on general research ethics have taken place within a psychological frame. It is generally agreed that active human participation of any sort necessitates adherence to basic ethical principles, often stated as confidentiality, informed consent, and a participant's right to withdraw themselves or their data from a study (APA, 2010; Kraut et al., 2004; National Research Council, 2014). Relevant funding bodies stipulate demonstration of adherence as do academic outlets for publication. Correspondingly, ethical guidelines are part of national research frameworks and are embedded in the research policy of academic institutions.

Bodies concerned with the development of online ethical guidelines have focussed mostly on issues of informed consent as in the case of accessing private, or semi-private, user information through internet services without explicit user permission (British Psychological Society, 2013; National Research Council, 2014; see also Moreno et al., 2013, for a discussion of informed consent in the online public vs. private domain). In other words, the focus is mostly on the process of data collection and on the context in which this happens, less on issues arising from particular qualities of the data obtained. While the former are comparatively well-known challenges, the latter require a deeper level of analysis. We argue that the amount and complexity of information readily obtainable from SNS, in collaboration with these services or with individual service users, is unprecedented and far surpasses the circumscribed data sets that are generated for most studies published in psychology. This generates a new need to understand what digital data exactly researchers are obtaining and handling.

At first, the sheer amount of information available through internet services is impressive. Findings reported on the basis of larger and richer data sets are often seen as more convincing, and more 'valid', by reviewers, editors, auditors, the media, and wider audiences. While larger data sets carry more statistical power, it is important to be aware of the standards set by computational disciplines such as Computer Science on this dimension. Behavioural track data on hundreds of thousands of SNS users are no rarity and sometimes make small effect sizes eminently publishable (e.g., Bond, Fariss, Jones, Kramer, Marlow, Settle, et al., 2012; Kramer, Guillory, & Hancock, 2014). While concerns have been raised regarding the interpretations based on such data sets (Ruths & Pfeffer, 2014), the ethical dimensions of having large amounts of information at the disposal of researchers have rarely been discussed (see Gleibs, 2014, for an overview).

A particular challenge lies with the anonymization of SNS data. Simply re-labelling nodes, the listed contacts, in a network does not remove the structural information that the location and connectedness of a particular node within the network carry. This first became apparent in the attempts to make a comparatively small Facebook data set on 1640 users publicly available (Kaufman et al., Zimmer, 2010). Since all users shared a common location and could be placed within a particular time window, individual users were at risk of becoming recognisable despite removing their names. This version of the anonymization problem, more technically speaking, can be conceptualised as a re-matching of a list of known names to a set of nodes. In a more general form, works by Narayanan and Shmatikov (2009) and Hay et al. (2008) on re-identification shows how anonymized nodes in large networks of, say, 10^6 SNS users become identifiable if sufficient independent information is held for particular users. Our own approach presents a different anonymization problem based on the type of information that psychologists are likely to collect in network studies. We argue that SNS users become identifiable by necessity based on comparatively few pieces of information on their ego-net, easily collected in small-sample research.

The non-anonymity of SNS data can seriously compromise adherence to the fundamental principles of research ethics, in particular where confidentiality is concerned. A participant's right to withdraw data from a study is not challenging in the case of ego-nets. Either data are in the public domain and can be obtained through observational methods, or data collection necessitates some form of interaction between researcher and participant, and procedures for data withdrawal can be put in place. Likewise, Moreno et al. (2013) provide a detailed analysis of issues surrounding informed consent based on the stated purpose of web services, their terms and conditions of use, and their privacy policies. Confidentiality, as the last and potentially most profound principle, can only be maintained through careful consideration of issues surrounding data archiving and data access. If, as we argue, some ego-

net data have to be treated as a unique participant identifier in themselves, then researcher awareness is crucial when handling such information. To date, online data do not hold a specific recognised status, in contrast to, for example, protected health information (PHI) under US law. This lack of status necessitates a closer look at how long-term confidentiality can safely be maintained when network data are part of a research project.

With these points in mind, we will next outline the uniqueness of ego-nets in a step-by-step manner, following in the footprints of a researcher from the social-behavioural sciences, rather than as a stringent mathematical treatment. This, we hope, will help to make the complexity of ego-net data more appreciated in fields that focus less on computational issues.

The uniqueness problem of SNS data

Our basic argument is that a comparatively small number of alters can be interconnected in so many different ways that each specific ego-net is highly likely to be unique to an individual within the world population. The information contained in the network structure takes on the role of a sociometric marker. In the context of the present work, we will call this the uniqueness problem of social network data. The uniqueness problem echoes classical debates in psychology on the diagnostic values of stereotypical attributes and personality traits. How a set of social attributes and the variability on these attributes influence person perception lies at the heart of memory models of social cognition (e.g., Linville, Fischer, & Salovey, 1989; Linville & Fischer, 1993). Given sufficient variability on a small number of social attributes, a perceiver will be able to differentiate between two members of the same group with a sufficiently high likelihood. Following this logic, it should be possible to determine how many attributes at which level of variability are necessary to, say, allow for an individual profile for each human being on the planet. Similarly, trait-based personality models were

introduced to psychology together with the claim that they allowed for nuanced profiles to emerge for individuals at a comparatively low number of traits (Cattell, 1943; 1949). We argue that information on networks, even where composed from very few attributes, is far more precise and therefore of much higher diagnostic value.

For our analyses, we will follow the traditional formalisation of a social network as a graph – an abstract set of nodes (online contacts, people) and edges (social ties, any other form of connection). The computation of graph properties lies at the heart of Social Network Analysis (Scott, 2013) and other applications of graph theory (Gross & Yellen, 2005). It does not matter, in the present context, whether SNS data are conceptualised as ego-nets or whole networks. In the following examples, an investigation of the alter nodes, without considering the edges involving ego, is sufficient to outline the uniqueness problem. Alter nodes and edges provide a simplified graph model of the ego-net, and the only important thing to keep in mind is that such a graph will not contain any truly isolated alters. Every alter node is connected with any other through the ego at the very least. A real-life example is provided by Figure 1 which shows a Facebook network of 138 alters arranged in such a way as to demonstrate clustering. Every node in this graph visualisation shares an invisible tie with the anonymous ego of this particular Facebook account.

This assumption makes it possible to apply some standard techniques for graph enumeration to ego-nets. Graphs, of course, are notorious for creating mathematical problems of exceeding computational complexities (Cooke, 2011; Köbler, Schöning, & Toran, 1993). The challenges of simple graphs have become known to wider audiences through the traveling salesman (or salesperson) problem (Cooke, 2011), but most researchers with a background in psychology will have come close to similar problems in their statistics training. Combinatorics, as used to illustrate statistical sampling problems, have a lot to do with graph enumeration which we will use in the following for a quantified demonstration of

how social network data are bound to form sociometric markers, even in fully anonymized format and with sparse information.

Scenario 1 – Networks containing unique nodes

Consider a network where all members are known by name. This would be equivalent to a graph consisting of unique, identifiable nodes. Assuming only one type of social tie and the fact that the network is held together by ego, how many different ways are there to distribute edges over all possible alter-alter pairs? Figure 2 illustrates this problem in the form of graphs and corresponding tie matrices.

The number of all possible graphs at a given network size n is determined directly by the number of edges in the graph ($k = (n^2 - n)/2$) and is $u = 2^k$. A small clique of 10 identifiable alters could be connected in $u = 2^{45} = 35{,}184{,}372{,}088{,}832$ different ways, and a Facebook network of 150 alters (which may or may not be the average network size of a Western adult user; see Dunbar, Arnaboldi, Conti, & Passarella, 2015; Duggan, Lenhart, Lampe, & Ellison, 2015) could manifest itself in $2^{11{,}175}$ different combinations. If we take u as our measure of uniqueness, it would be safe to say that the way in which a specific set of online contacts is structured is highly unique to this set, in particular when u gets compared to the global population, currently at below 2^{33}.

It is important to note that 2^k provides an upper limit to this sort of complexity. In reality, online social networks are much more likely to assume some structures than others, with significant and characteristic clustering observable (Buglass et al., 2016). In fact, it has been argued that the distribution of online network indices will be restricted to certain boundaries given human cognitive capacity limits (Dunbar et al., 2015). However, even discarding more than 99.99% of all possible combinations as humanly irrelevant would leave $2^{11{,}175}$ divided by, say, $2^{14} = 2^{11{,}161}$ combinations for the average Facebook network size.

While a truthfully labelled graph is rarely of relevance to researchers after data have been obtained, the combinatorics rely on the non-interchangeability of nodes, not on the use of true names of contacts. Replacing real names with any other set of labels will not change the uniqueness score of an ego-net. A response to this issue could be to fully anonymize the graph, but this will only change the uniqueness property of the data if anonymization is irreversible and all information on a node's identity is removed from the data set. This is the scenario to be considered next.

Scenario 2 – Networks containing interchangeable nodes

The anonymization of network data means that the examples in Figure 2 no longer apply. In Figure 2, all nodes and every edge (A-E, C-B, D-E and so forth) are unique. Anonymization renders both nodes and edges interchangeable. This does not mean that graphs lose all their characteristics. What is left in terms of structural information can be expressed as the number of non-isomorphic graphs, given a particular network size. Two graphs are non-isomorphic if they cannot be rendered identical through standard operations on their topology such as rotating, reflecting, stretching, and so forth. The number of non-isomorphic structures depends, again, on the number of nodes and edges. As a result, u will be dramatically smaller than 2^k, but the numbers are still growing exponentially with network size. Figure 3 picks out three non-isomorphic structures for a small network of four nodes and their isomorphic variants, a-d.

For anonymized graphs, u can be calculated using Polya's Theory of Counting or Polya-Redfield counting (Gross & Yellen, 2005). This way of counting imposes non-trivial computational demands as network size grows, but it can be used here for small numbers to provide a direct comparison with non-anonymized graphs. Polya counting can be stated as a colouring problem: How many different ways are there to colour in all edges of a graph?

Again, different here means discounting for any rotating, stretching, or moving around of nodes. For Polya counting, a cycle index polynomial is needed, a summative expression of all the different ways in which permutations can happen within the graph, and this can then be expanded by inserting the desired number of colours and counting the resulting terms when expanding the polynomial expression. At the outset, two colours can be used to represent the presence and absence of ties, in line with our model of an ego-net. This way, the process of counting will cover all graphs of a given size at all possible densities. In other words, the two-colour counting task yields the number of non-isomorphic graphs after complete anonymization of nodes. The use of two colours is illustrated in Figure 3 by the two different types of edges in the graphs. Using the appropriate cycle index polynomial for graphs of size $n = 4$, the number of non-isomorphic graphs is $u = 11$, as compared to $2^6 = 64$ for non-anonymized nodes; for $n = 5$ $u = 35$, as compared to $2^{10} = 1024$.

There is still the possibility that larger graphs, of average Facebook size, may compromise a user's privacy in the sense that the network structure itself, even without any further ego or alter attributes, is sufficiently unique to be regarded as a user's digital signature. Comparing the structures of several online ego-nets of average size against each other, however, requires considerable resources in terms of data access and computation.[2] For smaller isolated segments of an ego-net, the example figures suggest that the uniqueness problem is substantially reduced.

Scenario 3 – Networks containing interchangeable nodes and their categorical information

While anonymization brings the uniqueness problem down to a more manageable scale, it is important to consider that data on alter-alter ties may also include additional information on alters, for example, basic demographics. Such information would be relevant in virtually any

piece of psychological research. In fact, there are few reasons for psychologists to deliver a structural network analysis without any additional information on the egos and alters involved. Gender composition or age groups within a social network, emotional closeness to alters, duration of relationships, family or friendship bonds – there are numerous examples that would strike any researcher as potentially useful information.

Any piece of information that is available on alters, however, exacerbates the uniqueness problem to considerable degrees since it requires expanding the set of colours used in Polya counting. This is because with further alter attributes there are more than two types of ties to consider. Take sex of alters as an example. In its most condensed form, this information would necessitate three colours to designate three types of ties: no tie, same-sex tie, cross-sex tie. It would be more appropriate, however, to use four colours: no tie, female-female, male-male, and female-male, and there may be expansions of these classifications. Any alter variable with three categories would allow us to distinguish between seven types of ties, and so forth. More formally speaking, the number of required colours c using just one categorical variable is $c = (m^2 - m)/2 + m + 1$ where m is the number of categories. Suppose that we have more than just one alter attribute such as sex of alters plus alter-alter kinship. At this very basic level, we would need at least five colours to characterise network ties: no tie, same-sex related, same-sex unrelated, cross-sex related, cross-sex unrelated. As is shown in Figure 4, alter attributes bring back variety to the network graphs thereby rendering isomorphic structures distinguishable again.

It is in a way obvious that the complexity of anonymized graphs will grow exponentially the more information is available on alters and their social ties. Just how complex is demonstrated in Figure 5 that illustrates the outcome of some Polya counting exercises manageable to researchers outside of computation-heavy disciplines.

Taking the numbers from non-anonymized graphs as a comparison, it becomes obvious that the combination of alter-alter ties with even the simplest alter attributes elevates levels of uniqueness substantially. With just one dichotomous attribute such as gender, the possible combinations grow beyond the 2^k that hold for non-anonymized networks. While there is no clear-cut algorithm for bigger network sizes within easy reach, it is clear that 2^k will provide a lower boundary for *u* at best whenever any alter attributes are part of the data structure. This means that a small ego-net or a cluster within a larger network comprising of no more than 10 alters, for example, can bring *u* to levels that would allow for identifying every single user in the global population. The likelihood of any two users to produce the same network, or even network clusters, is getting smaller and smaller the more alter attributes are added to the investigation and becomes negligible for what we would deem small networks (e.g., 20 nodes) and few dichotomised attributes (e.g., 3).

As with non-anonymized networks, it needs to be stressed that the vast majority of possible graphs would stand very little chance of manifesting themselves in the field. What would be the likelihood of encountering an all-male network of 150 nodes with maximum density, i.e., with all possible interconnections in place? But once again, even discarding 99.99% of all possible combinations would reduce their numbers by a mere factor of 10,000, not nearly enough to get the magnitude of *u*, even for small ego-nets, down into the regions of the world population.

Implications and Recommendations

The purpose of outlining the uniqueness problem of online network data in the present work is to draw attention to particular characteristics of data that are of outstanding value to psychological research, yet have no strong history within the discipline. By demonstrating why, how, and to which extent such data can act as sociometric markers, we intend to raise the profile of this data type within the discipline and to contribute to a timely ethical debate

on digital data management. While a lot of information to be found on social media can be gathered through observation of a public online space without raising any particular ethical concerns (Moreno et al., 2013), there is also information obtainable that presents a unique trace to individual users, even with anonymization measures in place. This potentially compromises the adherence to the ethical principle of confidentiality and has implication for data archiving and sharing.

The uniqueness problem would be a minor one at best in the offline world where there is no service provider controlling a digital data base of all the information of interest. Information on offline networks does not exist, until it is created and compiled by a researcher, shows no compatibility with information on others' networks, unless manually formatted in specific ways, is typically safeguarded by ego and likely to depend on self-reports and so forth – restrictions that do not hold for digitally managed data. The analogy would be to come across the fully sequenced genetic code of an individual, without a name tag, but with the certainty that the code is on file somewhere. Further, (a) the code can be visualised in such a way that it becomes accessible to human perception, (b) substantial parts of the code link with those of other individuals, likewise on file, and (c) developing ways of exploiting this information is currently prioritised by academics and non-academics.

Two main points of criticism can be raised against the uniqueness problem: these are changes in network structure over time and feasibility of computation. Online networks are not stable since users add and, far less frequently, delete contacts. This change affects structural properties of the network and means that a sociometric marker is bound to become invalid over time. While changes are most likely to occur on the periphery of an ego-net, among looser connections, established clusters of online friendship ties are much more likely to remain stable due to their additional offline maintenance as well as due to stabilising processes such as structural balancing (Binder, Howes, & Smart, 2012; Szell, Lambiottte, &

Thurner, 2010). Focusing on the more central clusters within an ego-net also increases the feasibility of computation. As our simulation shows, even a sub-part of an ego-net of about 20 alters can easily exhibit uniqueness to a problematic extent.

While researchers and practitioners in psychology are used to handling sensitive information, and this receives ample space in ethical protocols, it is doubtful that the sensitive nature of network data is easily recognised. To return to an earlier example of a study design that contains both sensitive clinical data constituting PHI and data on online networks: Reviewers, commentators and institutional research boards would traditionally devote much more thought on the former data type than the latter and focus on the dissociation of known identifiers from the PHI. In such a case, however, identification of individuals may happen through the network information instead of other identifiers in the wider data set. This needs to be considered when it comes to data archiving for future re-use and sharing within larger collaborations. In psychology, the recurrent use of data sets is so far mostly known to those working in the health and clinical domains where comprehensive longitudinal data sets are generated and shared (e.g., Grant et al.; 2004). Again, such information can differ crucially from online network data and requirements for appropriate data archiving will depend on the precise data type.

The detailed analysis of the uniqueness problem allows for the systematic derivation of a set of hierarchical recommendations for managing online network data. Taken together, these recommendations can be applied to a wide range of study formats and designs, from the computation of social network indices across a larger number of ego-nets to the illustration of detailed case studies of a limited number of networks.

1. Where possible treat the full information gathered on online ego-nets as raw data requiring confidentiality. In many cases, data archiving requirements and research

collaboration needs can be satisfied by the sharing and publishing of data sets that contain aggregate ego-net characteristics (such as density or clustering indices) together with other individual-level variables. This information does not typically allow for deanonymization. Unless other details on the online location of the ego-net are disclosed, e.g., the name of a Facebook group, uniqueness will be very low. This case is analogous to publishing network visualisations, another core area of Social Network Analysis. Visualisations come with drastic information loss for larger networks and do not normally allow for a reconstruction of the full network data set.

2. In case of a requirement to publish or share the actual network data, researchers should consider the approximate level of data uniqueness. If the mere structural information is to be considered with fully interchangeable nodes, as outlined in Scenario 2, uniqueness will be very low and deemed unproblematic. Indeed, these are the data sets that are currently accessible to researchers for further analyses, which will inevitably be restricted to structural features (e.g., Dunbar, Arnaboldi, Conti, & Passarella, 2015). Note that with growing network size and stability, user identifiability becomes more feasible. The uniqueness problem does not fully disappear. Since online structures are likely to undergo at least some change over time, however, these concerns can be alleviated in most cases.

3. Where full network data are to be published or shared beyond mere structure, consideration of uniqueness levels should be taken into account more explicitly. Uniqueness will be highly sensitive to the number of additional variables recorded for alters, to a much higher degree than any amount of information known about the ego. This is outlined in Scenario 3. As Figure 5 illustrates, any additional type of alter-alter tie that can be defined raises uniqueness by about the magnitude of ten. This means, only the most basic of demographic information, e.g., alter gender, alter age band, can safely be stored with the structural information, and only so in a very small number of variables, e.g., not exceeding

three. For as long as contextual information on the ego does not provide further clues to the online network location, these small amounts of alter-alter information are unlikely to cause problems.

4. The sharing and publication of network data containing more and richer information on alter-alter ties is the case that requires most attention from researchers. In such cases, routine anonymization efforts are likely to be ineffective due to high levels of uniqueness. The threshold is very low. In line with our estimates so far, once three basic demographic variables are reached, a network of typical size, and even the larger clusters in it, can be regarded as a highly personalized sociometric marker. Several options can provide a remedy (see Corti, Van den Eynden, Bishop, & Woollard, 2014, for a general discussion of good practice in data archiving). Access to the data may be controlled such that third parties have to register and request access and agree to terms and conditions of data use. Alternatively, or even in addition to access restrictions, variables can be split up into separate data sets with no immediate correspondence between cases. This would prevent third parties from reconstructing the full network information. Similarly, it may be useful to remove some of the alter attributes from the full data set, which would strongly decrease uniqueness, and to report information on these as statistical aggregates.

It is hoped that these recommendations will help researchers to avoid any pitfalls when dealing with network data. At the same time, there are considerable research opportunities that arise from a study of online networks. The uniqueness problem stems from enormous data richness and the high diagnostic potential associated with social networks. These network properties, from our perspective, are still awaiting productive exploitation in psychology. Data uniqueness should therefore not only be seen as challenging, but should also provide further inspiration and facilitation of the research process when it comes to study

design, obtaining ethical approval from institutional research boards and the writing of grant proposals, in particular when it comes to demonstrating competent data management.

We conclude by noting that the uniqueness problem touches on issues of general online security and the rules governing public and semi-public cyberspace. Cybersafety and cybercrime provide the main context for studies on online anonymization problems (Hay et al., 2008; Narayanan & Shmatikov, 2009), and in this context ethical considerations for the purposes of research appear to be of secondary interest. User privacy and identifiability, however, are quickly turning into everyday concerns and are no longer confined to particular user groups as the recent case of FindFace demonstrates (Frankle, 2016). FindFace is an application for automated facial recognition that can be used to match any publicly obtained image to any other online source. In analogy to the uniqueness problem, this is another instance of identifiability of individuals within a global user population based on a specific marker, here facial features, by digital means. All such scenarios come with novel opportunities and challenges, and an awareness of those among psychologists is important to ensure the discipline's currency.

References

APA (American Psychological Association) (2010). *Ethical Principles of Psychologists and Code of Conduct*. Washington, DC: APA.

Binder, J.F., Howes, A., & Smart, D. (2012). Harmony and tension on social network sites: Side-effects of increasing online interconnectivity. *Information, Communication & Society, 15*, 1279-1297. doi: 10.1080/1369118X.2011.648949

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489(7415)*, 295-298. doi: 10.1038/nature11421

boyd, D., & Ellison, N. B. (2008). Social network sites: definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*, 210-230. doi: 10.1111/j.1083-6101.2007.00393.x.

British Psychological Society (2013). *Ethics Guidelines for Internet-mediated Research.* Leicester: British Psychological Society.

Brooks, B., Hogan, B., Ellison, N., Lampe, C., & Vitak, J. (2014). Assessing structural correlates to social capital in Facebook ego networks. *Social Networks, 38*, 1-15. doi: 10.1016/j.socnet.2014.01.002

Brown, J., Broderick, A.J., & Lee, N. (2007). Word of mouth marketing within online communities: Conceptualizing the online social network. *Journal of Interactive Marketing, 21*, 2-20. doi: 10.1002/dir.20082

Buglass, S.L., Binder, J.F., Betts, L.R., Underwood, J.D. (2016). When 'friends' collide: Social heterogeneity and user vulnerability on social network sites. *Computers in Human Behavior, 54*, 62-72. doi: 10.1016/j.chb.2015.07.039

Cattell, R. B. (1943). The description of personality. I. Foundations of trait measurement. *Psychological Review, 50*, 559-594. doi: 10.1037/h0057276

Cattell, R. B. (1949). R p and other coefficients of pattern similarity. *Psychometrika, 14*, 279-298. doi: 10.1007/BF02289193

Cooke, W. (2011). *In pursuit of the traveling salesman: Mathematics at the limits of computation*. Princeton, NJ: Princeton University Press.

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and sharing research data*. London: Sage.

Crossley, N., Bellotti, E., Edwards, G., Everett, M.G., Koskinen, J., & Tranmer, M. (2015). *Social network analysis for ego-nets*. London: Sage.

De Laat, M., Lally, V., Lipponen, L., Simons, R. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. *Computer-Supported Collaborative Learning, 2*, 87-103. doi: 10.1007/s11412-007-9006-4

Duggan, M., Lenhart, A., Lampe, C., & Ellison, N.B. (2015). *Parents and social media.* Washington, DC: Pew Research Center.

Dunbar, R. I. M., Arnaboldi, V., Conti, M., & Passarella, A. (2015). The structure of online social networks mirrors those in the offline world. *Social Networks, 43*, 39-47. doi: 10.1016/j.socnet.2015.04.005

Fiske, S.T., & Hauser, R.M. (2014). Protecting human research participants in the age of big data. *Proceedings of the National Academy of Sciences, 111*, 13675 – 13676. doi: 10.1073/pnas.1414626111

Frankle, J. (2016, May 23). *How Russia's new facial recognition app could end anonymity.* The Atlantic. Retrieved from http://www.theatlantic.com

Gleibs, I. H. (2014). Turning virtual public spaces into laboratories: Thoughts on conducting online field studies using social network sites. *Analyses of Social Issues and Public Policy, 14*, 352-370. doi: 10.1111/asap.12036

Grant, B. F., Stinson, F. S., Dawson, D. A., Chou, S. P., Dufour, M. C., Compton, W., ... Kaplan, K. (2004). Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: Results from the national epidemiologic survey on alcohol and related conditions. *Archives of General Psychiatry, 61*, 807-816. doi: 10.1001/archpsyc.61.8.807

Gross, J.L, & Yellen, J. (2005). *Graph theory and its applications*. Boca Raton, FL: CRC Press.

Hay, M., Miklau, G., Jensen, D., Towsley, D., & Weis, P. (2008). Resisting structural re-identification in anonymized social networks. *Proceedings of the VLDB Endowment, 1*(1), 102-114. doi: 10.14778/1453856.1453873

Köbler, J., Schöning, U., & Toran, J. (1993). *The graph isomorphism problem: Its structural complexity*. Boston, MA: Birkhäuser Boston.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*, 8788-8790. doi: 10.1073/pnas.1320040111

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: report of Board of Scientific Affairs' Advisory Group on

the Conduct of Research on the Internet. *American Psychologist, 59*, 105 - 117. doi: 10.1037/0003-066X.59.2.105

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks, 30*, 330-342. doi: 10.1016/j.socnet.2008.07.002

Linville, P. W., & Fischer, G. W. (1993). Exemplar and abstraction models of perceived group variability and stereotypicality. *Social Cognition, 11*, 92-125. doi: 10.1521/soco.1993.11.1.92

Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: empirical evidence and a computer simulation. *Journal of Personality and Social Psychology, 57*, 165-188. doi: 10.1037/0022-3514.57.2.165

Moreno, M. A., Goniu, N., Moreno, P. S., & Diekema, D. (2013). Ethics of social media research: common concerns and practical considerations. *Cyberpsychology, Behavior, and Social Networking, 16*, 708-713. doi: 10.1089/cyber.2012.0334

Narayanan, A., & Shmatikov, V. (2009). De-anonymizing social networks. *Proceedings of the 30th IEEE Symposium on Security and Privacy*, 173-187. doi: 10.1109/SP.2009.22

National Research Council (2014). *Proposed revisions to the common rule for the protection of human subjects in the behavioral and social sciences*. Washington, DC: National Academies Press.

Puschmann, C. & Bozdag, E. (2014). Staking out the unclear ethical terrain of online social experiments. *Internet Policy Review, 3*. doi: 10.14763/2014.4.338

Research at Facebook (Feb, 2016). *Three and a half degrees of separation*. Available at https://research.facebook.com/blog/three-and-a-half-degrees-of-separation/

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346(6213)*, 1063-1064. doi: 10.1126/science.346.6213.1063

Scott, J. (2013). *Social Network Analysis*. London: Sage.

Shah, D.V., Cappella, J.N., & Neuman, W.R. (2014). Big data, digital media, and computational social science: Possibilities and perils. *Annals of the American Academy of Political and Social Science, 659*, 6-13. doi: 10.1177/0002716215572084

Szell, M., Lambiotte, R., & Thurner, S. (2010). Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences, 107*, 13636-13641. doi: 10.1073/pnas.1004008107

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). *The anatomy of the facebook social graph*. Available at: http://arxiv.org/abs/1111.4503

Wilson, R.E., Gosling, S.D., & Graham, L.T. (2012). A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science, 7*, 203-220. doi: 10.1177/1745691612442904

Zimmer, M. (2010). “But the data is already public”: On the ethics of research in Facebook. *Ethics and Information Technology, 12*, 313-325. doi: 10.1007/s10676-010-9227-5

Footnotes

[1] Ease of access varies and is often subject to the current terms and conditions of particular web services. Research on Facebook ego-nets, for example, has been substantially complicated since a change to the application-programmer interface in 2015. Such fluctuations notwithstanding, online network information that is visible to users is also, in principle, usable for research purposes.

[2] In the world of online data, researchers are advised to think in the area of millions of users rather than at the scale of experiment-based samples. Full computation may not always be feasible. For very large networks stochastic methods have to be used, as in a recent analysis of the full Facebook graph (Research at Facebook, 2016).

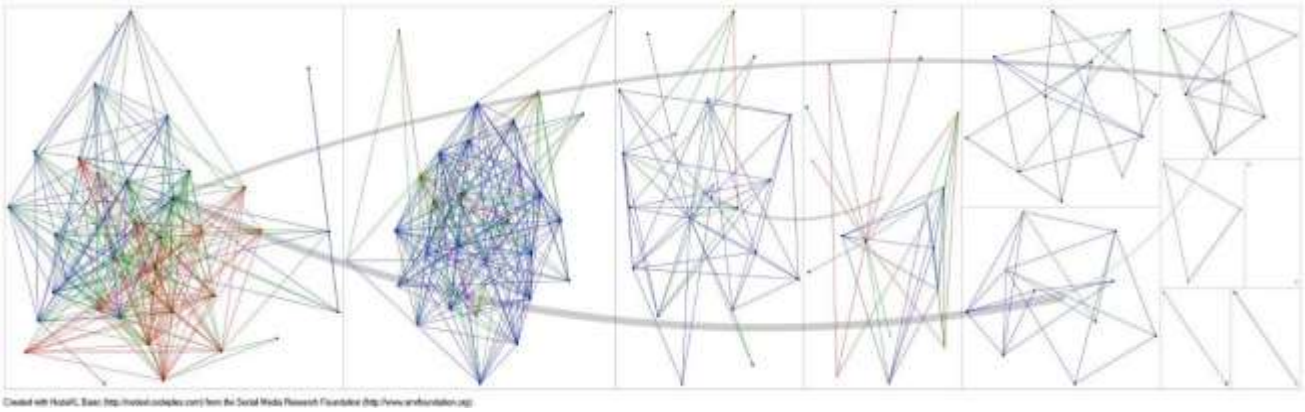
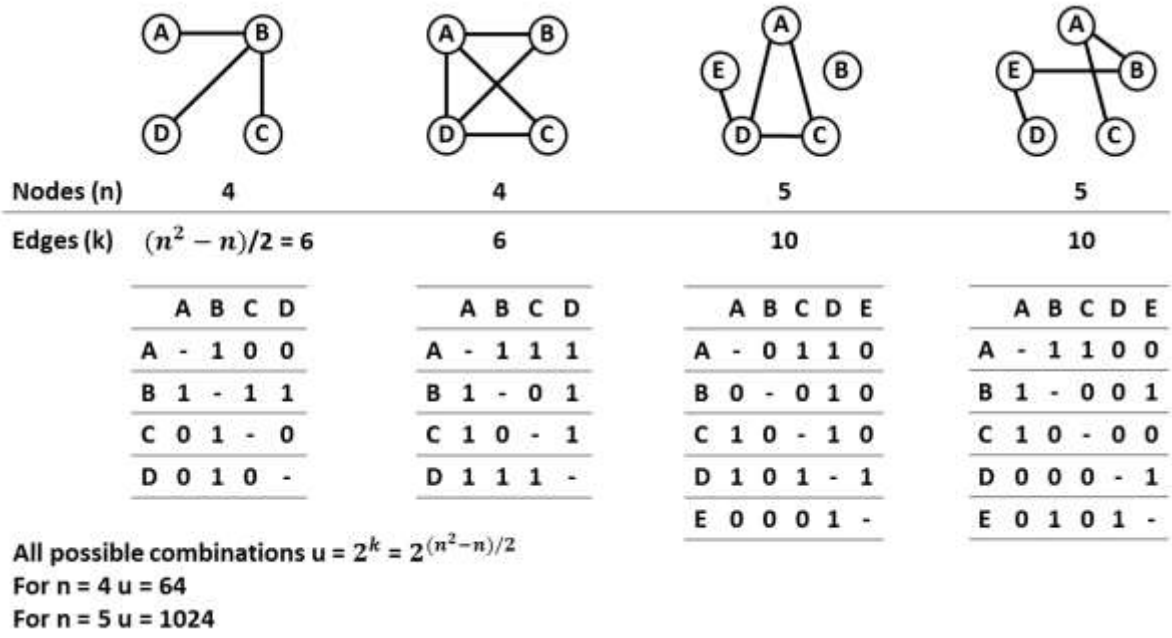


Figure 1. Example of a Facebook ego-net containing 138 nodes displayed to maximise distinctiveness of clusters. Edges in red represent female-female ties, blue male-male, green female-male. Long curved ties are those bridging clusters.



	A	B	C	D
A	-	1	0	0
B	1	-	1	1
C	0	1	-	0
D	0	1	0	-

	A	B	C	D
A	-	1	1	1
B	1	-	0	1
C	1	0	-	1
D	1	1	1	-

	A	B	C	D	E
A	-	0	1	1	0
B	0	-	0	1	0
C	1	0	-	1	0
D	1	0	1	-	1
E	0	0	0	1	-

	A	B	C	D	E
A	-	1	1	0	0
B	1	-	0	0	1
C	1	0	-	0	0
D	0	0	0	-	1
E	0	1	0	1	-

Figure 2. Graph combinations in non-anonymized networks. A-E represent individual contacts (alters) of the network owner (ego). All alters are therefore connected with each other at least through an indirect connection via ego, as is the case for node B in the third model from the left.

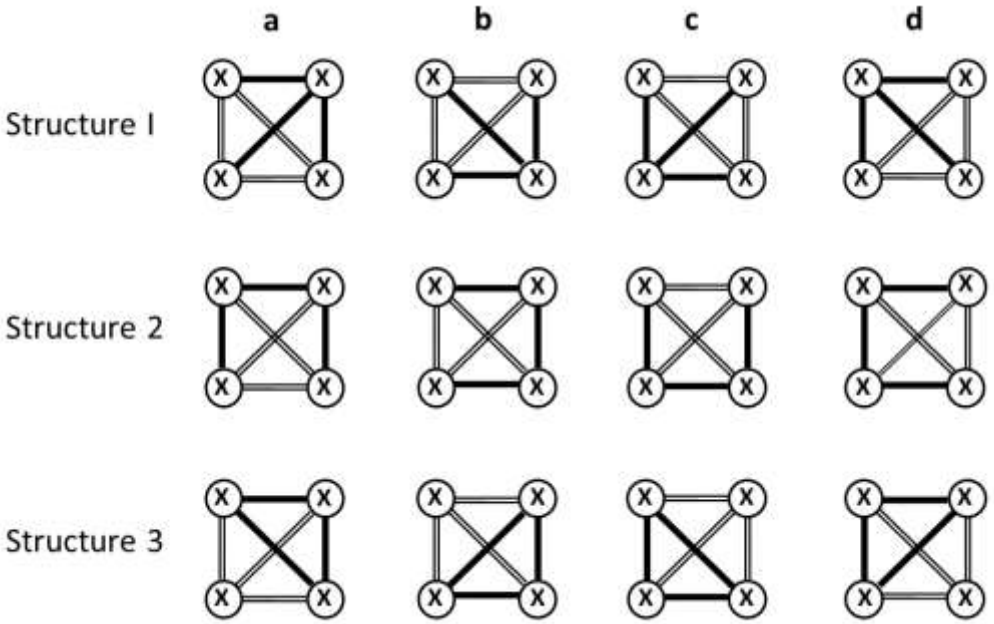


Figure 3. Anonymized graphs leading to isomorphic structures (a-d) that are indistinguishable from each other. Lines in black refer to actual ties, lines in white to the absence of ties.

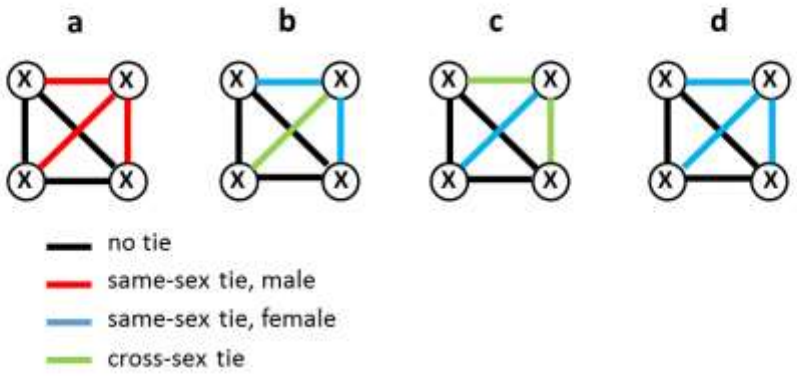


Figure 4. Anonymized graphs with minimum information on alter sex added. Previously indistinguishable structures a-d, as in Figure 3, are now distinguishable from each other.

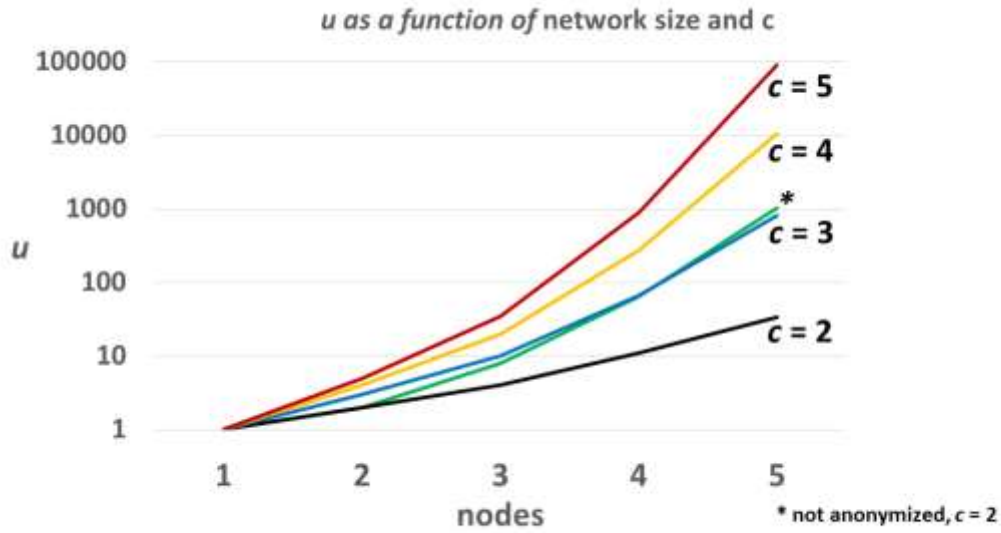


Figure 5. Uniqueness of a network depending on network size (number of nodes) and number of colours (c) needed to account for all types of alter-alter ties in the graph.