Soc. Netw. Anal. Min. (2014) 4:198
DOI 10.1007/s13278-014-0198-0

ORIGINAL ARTICLE

Building rich social network data: a schema to assist in designing, collecting and evaluating social network data

Eamonn O'Loughlin · Diane Payne

Received: 28 August 2013 / Revised: 15 April 2014 / Accepted: 3 May 2014 / Published online: 3 June 2014

Abstract Creating a social network dataset requires us to represent a set of empirical observations according to a specific conceptual understanding. This requires a number of design decisions for the conceptual framework, which is then implemented through a data structure. In this paper, we propose a standard schema to describe these decisions. A standard schema allows us to define the conceptual framework, structure, and content of a dataset. Social network datasets may contain many features. Beyond the definition of actors and relations, network data may include: actor or relation attributes; data for multiple observation periods (dynamic data); or parallel event data. The creation of a network dataset may also involve the application of specific boundary conditions, sampling approaches or may include missing data. Our proposed schema is designed to support a scientific approach to social network analysis by making these features and assumptions transparent and easy to communicate. We believe that this will facilitate researchers through the design, creation, communication, and evaluation of social network datasets. To develop this schema, we gathered and analysed the structure, content, and metadata of over 150 publicly available social network datasets drawing from multiple disciplines, including statistics, computer science, sociology, economics, and political science.

Keywords Social network data · Dynamic networks · Complex networks · Parallel network data · Data models and structure · Data collection

E. O'Loughlin · D. Payne (✉)
Dynamics Lab, UCD Geary Institute, University College Dublin, Belfield Campus, Dublin 4, Ireland
e-mail: diane.payne@ucd.ie

1 Introduction

We live in a world where physical location no longer places constraints on whom we interact with. In this world, the application of social network analysis (SNA) techniques to understand behaviour has become significantly more pervasive among many disciplines including sociologists, statisticians, computer scientists, psychologists, physicists, and economists. Social network analysis techniques are not new, but in the past, researchers were constrained to collecting small network datasets, often at great cost Wasserman and Faust (1994). For example, Kapferer (1972) sat in a tailor shop in Zambia for a significant portion of 1972 documenting every interaction that took place between employees of this shop during his stay. We can easily see that the personal and economic commitment required to collect this data resulted in few others collecting similar data. Furthermore, without a significant investment of resources this approach is not scalable beyond very small networks. In more recent times technological advances, and the resultant mass adoption of these advances, has reduced the cost associated with sourcing and collecting data relating to social networks. We have seen that the size, scope and complexity of analysed social network data have increased substantially; we are now collecting network datasets that are bigger and contain more features (Farrugia et al. 2011). However, collecting larger and more complex datasets has brought a different set of challenges to the fore. When compared to multi-dimensional data (e.g. time-series), the creation of a social network dataset involves a significantly larger number of methodological and design decisions (Giles et al. 2010). We also observe that for the same empirical phenomenon, the decisions taken can lead to the production of significantly different datasets. For example, we may

group interactions with time periods of different lengths, define actors or relations differently, or exclude specific actors or relations from the captured data. In addition, the theoretical framework that underlies these decisions determines whether or not a dataset remains suitable for particular types of analysis. Collecting richer network data means that in addition to asking questions about efficient ways to store, manipulate, and perform operations on data (i.e. implementation decisions), we also must consider approaches to capture and describe the *features* and *decisions* represented in the network data.

In parallel to the changes outlined above, the increase in researchers who work with social network datasets has, in some cases, led to approaches (to structuring, collecting, and describing network datasets) that are fragmented across different disciplines. This trend hampers the development of network science as one science. In particular, and of interest here, different approaches to describing the data collected can result in a research environment where sharing datasets, methods, and insights between different disciplines are unnecessarily difficult.[1]

Besides some recent examples (Hennig et al. 2012) research into and conversations about the representation and storage of network data have tended to concentrate on storage systems suitable for fast operations on simple data structures, rather than the design of data models or schema optimised to hold many features. For example, Newman (2010), in summarising storage options for network data, concentrates on efficient storage options (including adjacency list, heap, and tree data structures), yet does not consider in detail the representation of actor or edge attributes or consider at all dynamic network data or the storage of data metadata. Finding a standard mechanism to represent data in a way that puts a structure behind the description of the salient features is the focus of this paper. Rich network datasets (i.e. datasets with many features) can still be reduced to adjacency lists, matrix structures or heaps as a data pre-processing step, for those who have no use for the additional features.

To help in addressing the challenges discussed above, we propose a schema for describing the features of social network data.[2] A schema allows us to represent, *in a standard way,* the features of a particular object (in this case the object is a social network dataset). This schema allows the researcher to understand, discriminate, and compare different network datasets by reference to their features. There are three main areas of motivation driving the creation of this schema: (1) it can assist in framing or thinking about how to approach the data sourcing or collection; (2) it allows easy communication about the conceptual framework and features of social network datasets between researchers; (3) it allows others, including software developers and statistical programmers, to anticipate the type of data that will become available and develop tools or algorithms to fit this data. We developed this schema by preforming an analysis on a list of more than 150 publically available social network datasets compiled in early 2013. The datasets date range is from 1905 to 2012, and size from 34 nodes to in excess of 5 million nodes. Each of the datasets used is referenced below and a detailed description/metadata is accessible on an online application located on the University College Dublin Dynamics Lab website.[3]

The remainder of this paper is presented in two main parts. Firstly, the next section entitled *A Schema to Describe Social Network Data* describes and summarises the key components of our social network data schema, using existing datasets to demonstrate or illustrate particular features. Secondly, *Building Rich Social Network Data* outlines how to use the schema to support the designing, collection and communication of social network dataset in a number of different contexts.

2 A schema to describe social network data

The reason for having a schema is that it makes collecting, working with, sharing and evaluating social network data easier. A good schema will allow us to summarise in an intuitive way the different features a particular social network dataset has or does not have. Given that social network datasets often range from the very simple (small in size, with few features) to the very complex (large datasets and with many features), a good schema should work well in both contexts. That is, in all cases the benefits of using a schema should outweigh the costs which in this case is the time taken to categorise a new dataset. In this section, we iteratively describe our schema by describing the features of social network datasets that we have identified in existing datasets. In each case, we begin by describing the feature and then illustrate an example of this feature using an existing publicly available dataset.

[1] A tangible example of this is the myriad of different social network data storage file formats. There currently exist at least 20 standard social network storage formats (including gexf, gml, GraphML, UNINet DL, Pajek NET, csv, edgelist, etc.) and many non-standard storage approaches. This can make data difficult to access for those unfamiliar with the data format and hence reduce the discovery of these datasets by specific disciplines.

[2] Here, we use the term feature according to its standard interpretation: 'a distinctive attribute or aspect of something'.

[3] University College Dublin Dynamics Lab http://dl.ucd.ie.

2.1 Minimal social network dataset

Under the standard paradigm for social network analysis, social networks are represented as a graph. In the minimal form, and indeed the de facto necessary condition, a social network dataset must include both nodes and edges for a snapshot of time or single time period.

The first component of a minimal social network dataset is the network nodes. In a social network dataset, a node (sometimes called 'Vertex', 'Actor') represents an individual unit of action or autonomy. This unit can be people, firms, organisations, countries or positions. In the minimal graph there is only one type of node. Nodes are usually represented by a label (name) or integer (identification). Some examples of different types of nodes include those studies by Zheleva (2011) where nodes represent individual member accounts for the pet social networking site *Dogster.com*; Newman (2006) where nodes represent scientists who have contributed to the Physics journal *Condensed Matter;* Brozovsky and Petricek (2007) in which nodes represent customers of the Czech dating site Libimseti.cz, Read (1954) which specifies each node as one of a tribe in eastern central highlands of New Guinea, and finally Batagelj and Mrvar (2006) who use nodes to represent academic papers which discuss a specific topic in network science. We observe that for social networks the fundamental unit—the node—can represent many different things. It is therefore obvious that clearly specifying what nodes in a dataset represent is of fundamental importance.

The second component of the minimum social network dataset is the network edges (sometimes called 'network ties', 'ties', 'links', or 'connections'). Edges represent some relationship between nodes. Later we will see that edges may have associated attributes like direction or weight—but the minimum representation is that of a simple 'undirected' edge. An undirected edge is an edge where the relationship is mutual. For example, if edges indicate collaboration between two researchers, the relationship is mutual, i.e. it is not from one *source* node to another *destination* node. This relationship can take many forms—for example, Gjoka and Kurant (2010) use an edge to represent friendship on the site facebook.com; Seierstad and Opsahl (2011) use an edge to indicate that two directors sit on the same board; for Loomis et al. (1953) an edge represents 'frequent gatherings' between different families; edges in Cross and Parker (2004) mean that both individuals were in attendance at the same meeting; edges in the dataset collected by Batagelj et al. (2008) mean that two researchers verbally interacted with each other; edges in Palla et al. (2008) represent the relationships where two researchers co-authored a paper; edges in Sampson (1969) represent self-reported friendship relations by junior members of a monastic order. Similar to nodes, we observe that for social networks the fundamental relationship—the edge—can be represented in many different ways. Clearly, different representations can signal significantly different relationships. For example, an edge might signify a relationship through an electronic medium, which rarely (if ever) involves an offline interaction. There has been some debate about the quality of these networks for use in academic research, covered neatly in Boyd and Ellison (2007). At this point, it should be obvious that clearly specifying what an edge in a dataset refers to is of importance.

Social Network Dataset Schema
Dataset Name
Description of Node
Description of Edge

Fig. 1 Minimal social network schema

Even at this point, we see that even in the minimal representation of a social network there are fundamental design decisions. For this reason we believe that a useful schema must begin with a clear specification of these decisions. For both nodes and edges, a brief description usually suffices (Fig. 1).

2.1.1 Layering in complexity

The minimal social network dataset contains the necessary features that must exist for the data to be a social network dataset. From this starting point, there are many different ways to add additional detail, complexity or scope into a dataset. We describe each of these below, and further outline them, using examples, in the next section.

2.1.1.1 Number of node types As we have seen, the minimal social network has one type of node. Collected network datasets, however, may have more than one type of node. In these cases, nodes represent different types of individuals, roles (e.g. authority positions), groups or organisations—for example, Stehlé et al. (2011) includes both a node type of 'teacher' and a node type of 'student'; Auer et al. (2007) include both a node type 'Music Artist' and a node type 'Record Label'. In both examples, the inclusion of additional node types adds richness to the dataset. A social network dataset with more than one type of node is a 'multimodal' dataset. Across the datasets that we have observed, it is not uncommon to see 2-mode, 3-mode, and 4-mode networks. In addition, if a network that has two different node types, and in this network edges can only exist between nodes of different types (and hence

Social Network Dataset Schema
Is the network dataset multimodal?
Is the network dataset bipartite?

Fig. 2 Schema: node type

never between nodes of the same type), we say that this network is multipartite. For example, for Auer et al. (2007) an edge represents 'Artist has a contract with music label' (an "artist ← → artist" or "music label ← → music label" direct edge cannot exist). In this example, there are two node types; the network is described as bipartite, as there are two such relationships. Across all the public datasets that we have observed, we have not observed any networks that are multipartite with a mode greater than two—so in order to maintain a parsimonious schema we include only the network characteristic 'bipartite' (Fig. 2).

2.1.1.2 Number of edge types and edge attributes The minimal social network has one type of edge. Datasets may include multiple types of edges. Social networks with edges of multiple types are called multiplex networks, and each edge represents a different type of relationship. For example, in Freeman and Freeman (1980) nodes represent academic researchers and there are two types of edges: (1) 'self-reported acquaintance with another researcher' and (2) 'number of messages sent during the period to this researcher'. Similarly in Lazega (2001), which is a network of lawyers at the same firm, there are three types of edge including 'co-worker edge', 'friendship edge', and 'advice edge'.

Irrespective of network type, edges may be described or categorised in further detail using attributes. These attributes are often used to describe specific semantics of the relationship. One such attribute is the 'direction' of the relationship. An edge may be directed from one (source) node to another (destination) node. In Prosper (2010) edges represent financial support given from an individual to a charity. We can see that in order for this type of edge to be created, the originating node is the creator of the relationship. Edges may also be ascribed a weight to indicate the strength or magnitude of the relationship. In the dataset collected by Opsahl and Panzarasa (2009) which is a network of users of a social networking site, edges represent communication between these users. In this network there is an additional edge attribute of weight, which is used to represent the frequency of communication (number of messages sent) (Fig. 3).

Another attribute commonly added to network edges is an edge sign. A signed edge is an edge with a value indicating that the relationship is either positive or

Social Network Dataset Schema
Is the network multiplex?
Network Edges
Undirected
Directed
Weighted
Signed (Positive / Negative)

Fig. 3 Schema: edge type and attributes

negative. The dataset used by Maniu et al. (2011) is a network where nodes are editors on the English version of Wikipedia and the edges represent interactions between the users. On Wikipedia, editors are able to restore or vote for another user's edit (a positive interaction) or revert or vote against this edit (a negative interaction). In the social network dataset, these interactions have been aggregated to an edge type of interaction with a sign indicating if this interaction is positive or negative.

2.1.1.3 Node attributes and known communities Network datasets may contain additional attributes that describe individual node attributes. These attributes may be fixed (e.g. an individual's gender) or variable (e.g. an individual's academic course of study or current geographic location). An example is the dataset collected by Newman (2006) which consists of a communication network between researchers and where the actor node attributes include 'primary discipline' and 'number of times they've been cited'. This attribute data adds an additional dimension to the actor, which could be an important component for the explanation of the structure of the network (Fig. 4).

In some cases, the node attributes indicate membership of an organisation or some other group affiliation. These attributes are often thought to make nodes with shared values more likely to have ties to each other. This membership attribute is a special type of attribute, and is often described as 'Community Membership'. If the community attributes that are known 'a priori' (i.e. before the network edges are captured) then they are sometimes called 'ground truth' communities. For example, in the social network dataset of Twitter communication between political blogs described by Adamic and Glance (2005), an attribute 'Liberal/Conservative' identifies actors as belonging to a community based upon their political affiliation. Similarly, the dataset of United Kingdom MPs (Greene and Cunningham 2013) contains the political

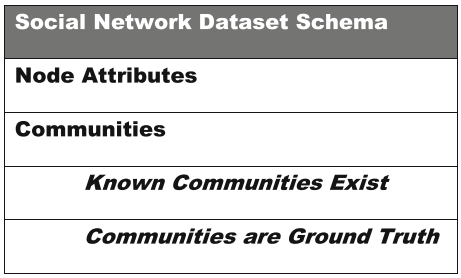

Fig. 4 Schema: including edge type and attributes

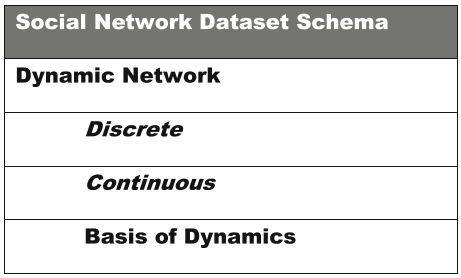

Fig. 5 Schema: dynamic data

party that they belong to as an attribute; this attribute is a ground truth community (as independent of the edges we could still observe the community).

2.1.1.4 Dynamic network data We have exclusively discussed datasets with no temporal component or dimension. Up to this point, we've assumed (implicitly) that each dataset described a snapshot of a network at a certain point in time. Such network datasets are called *static networks* and are by far the most common network dataset. However, datasets with a temporal dimension (called *dynamic network* datasets) have been growing in use. For example, most of the dynamic network datasets identified in our analysis are dated from the year 2000 and onwards. Providing a view of the network over multiple time periods immediately enables observation of the evolution of specific aspects of the network. Dynamic networks may take many forms. For example, the network may include dynamic edges (edges that are present in one time period, but not another) or dynamic nodes (where the node existence or node attributes change from one period to another). For an example of the former, we look at the Norwegian shared directorships dataset sourced by Seierstad and Opsahl (2011). A shared directorship is when two directors sit on two different company boards simultaneously. In this case, the temporal component is the change in the presence of edges between different directors as they are recruited to (or depart from) the same board. Over time the edges can be seen to exist/disappear between two directors. An example of the latter is the students' dataset (Van de Bunt et al. 1999; Van Duijn et al. 2003), which captures the smoking behaviour of students over several time periods. As the period changes, the student's attribute of smoking preference may also change (Fig. 5).

In dynamic network datasets temporal data may be represented using discrete/fixed time periods with a set number of equal distance observation periods/snapshots, or the dynamics may be captured through continuous time periods or event-driven time. In continuous or event-driven time period social network datasets, each edge is encoded with a date stamp to indicate the time/date that the edge was created, or a data range (indicating the date window inside which the edges are active).

2.1.1.5 Parallel data The datasets we have observed to this point include only attributes about nodes and edges within the network. However, in conjunction with the collection of social network data some researchers have also collected related data that is in the same context as the social network (i.e. either an antecedent or outcome of the network data). Often this is a descriptive list of events which occurred *before, during,* or *after* the collection of the network data—but these events are related to (or involve) actors that are within the scope of this network data. Collecting parallel data is often this first step is towards analysis investigating the relationship between events and network structure. A simple example is Kapferer (1972), which contains dynamic data on interactions between employees in a tailor shop and also includes parallel data specifying the date of a strike by the employees. Another example is the work by Zhao et al. (2006) which includes a dynamic network of known terrorists and their relationships—while also providing a parallel dataset of the dates and location of terrorist attacks. A final example is that of Leskovec et al. (2010), which contains directed edges of votes between Wikipedia users over successive elections (Fig. 6).

2.1.1.6 Data metadata The final component of the social network schema is not like the others, which listed specific attributes or features about the dataset—instead *data metadata* refers to decisions taken when collecting/creating the dataset. The first of these is the boundary conditions applied. Boundary condition refers to the mechanism used to include (and more importantly exclude) nodes or edges from the network. Nodes not present in the network may be a significant determinant of what is going on in the network, so if the logic used to define the boundary condition is unclear then the conclusions drawn may be bogus—specifying the boundary conditions is therefore important. A common boundary condition is to include only edges directly connected to some specified actor—these have

Social Network Dataset Schema
Parallel Data Description

Fig. 6 Schema: parallel data

been called 'Ego Networks'. The boundary condition in such an example would be that the network 'only includes network ties that are directly to/from this ego'. Other examples of boundary conditions are given by the Van Duijn students dataset (Van de Bunt et al. 1999) includes only the freshman cohort majoring in Sociology in the University of Groningen in 1996–1997; and the network of Scottish corporate interlocks (De Nooy et al. 2005) includes only the 108 largest firms for the period selected. In the same vein, the data in a social network dataset may have been extracted from a larger electronic or archival system. For example, each of the datasets discussed in Greene and Cunningham (2013) varies in terms of the set of actors included (Rugby Players and Premier League Players) and this reflects the boundary conditions applied in each case. However, these datasets were originally extracted from one large Twitter dataset and this knowledge or awareness, that further data exists, may be useful to help with future analysis of the dataset (or to explain findings which seem curious) (Fig. 7).

Similar to boundary conditions, and for reasons outlined in the introduction, it is important to identify missing data and (perhaps) more importantly if that missing data has a known pattern. Missing data can have a large impact on conclusions (Kossinets 2006), so clearly specifying patterns is essential. For example, the network outlined by Van de Bunt et al. (1999), which contains survey responses from students, only includes responses from students who have participated in/completed all seven surveys. Some students were excluded, yet the reason for this research design decision is not clear (e.g. Had they left the university? Refused to participate?) and without this information there is some uncertainty placed on the reliability of studies based on this dataset.

Unfortunately due to the nature of network datasets this is an unavoidable situation, and clearly communicating the method of exclusion (as the authors did) helps in constructing analyses. Another example is the dataset collected by Isella et al. (2011), which contains data relating to interactions between attendees at a conference. The data are available for 110 people collected over approximately 2.5 days where this data are only a subset of attendees and where only those who chose to participate in the network survey are included. Knowing this is important particularly if those who chose to participate shared some attribute

Social Network Dataset Schema
Data Metadata
Boundary conditions
Sampled from larger dataset?
Sampling approach documented
Missing data exists
Missing data pattern understood

Fig. 7 Schema: data metadata

(e.g. perhaps all drawn from same academic sub-community represented at the conference).

2.1.1.7 Full schema Having stepped through the different components of the schema, we summarise the full schema in the Fig. 8. The objective of this schema is to allow researchers to (1) assist in framing or thinking about how to approach the problem of data sourcing/collection; (2) facilitate easy communication about required or desirable features of social networks datasets between researchers; and (3) share with others, including software developers and statistical programmers, the semantic structure of the data that will become available so that they can develop products or algorithms that fit this data. Full schema details for all of the dataset references discussed in this paper is available on the University College Dublin Dynamics Lab website at http://dl.ucd.ie.

3 Building rich social network datasets

In order to source a social network dataset, we can choose one of three possible routes: (1) collect the data through direct observation or survey; (2) retrieve the data by taking a subset of data from an existing electronic or archival system; or (3) search for and assess the quality of an existing (available) social network dataset.

3.1 Collecting data through direct observation or survey

Collecting through direct observation or survey is traditionally one of the popular forms of data collection. The arrival of new technologies including phone sensor

Social Network Dataset Schema
Social Network Dataset Name
A node represents
An edge represents
Network is multimodal
Network is bipartite
Network is multiplex
Network Edges:
Undirected
Directed
Weighted
Signed (Positive / Negative)
Node Attributes
Communities:
Known Communities Exist
Communities are Ground Truth
Dynamic Network:
Discrete
Continuous
Basis of Dynamics
Parallel Data
Data Metadata:
Boundary conditions
Sampled from larger dataset?
Sampling approach documented
Missing data exists
Missing data pattern understood

Fig. 8 Full social network dataset schema

technologies or even online survey systems, make the collection of large original datasets easier. However, when conducting surveys, the time required can influence the number of legitimate responses. Whatever method of data collection is used, it is important to understand how best to optimise the process. We suggest that systematically stepping through the components of our schema to decide upon key design decisions may be productive. That is, start by describing what a node will be, what an edge will be and work through the more complex parts of the schema indicating if a feature will be present in the collected dataset (and later how this data will be collected). The schema we have documented was built by observing design decisions taken in a large number of different social network datasets. Using this schema to anticipate the structural questions others may ask, helps in making intelligent decisions about the (for example) the type of dynamic data that will be captured or how the boundary conditions will be specified.

In this context, the primary function, and best way to get advantage out of the schema is to use it as part of a structured exercise that formalises or makes explicit your project's data collection strategy and (hence) design decisions, while also (if necessary) assisting in communicating or sharing those design decisions with others working on the project. When it comes to describing your data collection approach, a populated schema will make this process less onerous and ensure that no aspects are missed.

3.2 Extracting data from existing electronic system

We observed this as the fastest growing method of creating social network datasets for research. Often as researchers we are extracting data from a system owned by a *for-profit* organisation. A for-profit organisation generally has two primary concerns: (1) protecting the privacy of their users or customers; and (2) ensuring sensitive business data (i.e. data that would assist competitors or data where public disclosure would be detrimental to their business) remain private (Narayanan and Shmatikov 2009). The process by which these data are sourced often begins with a discussion with an individual employed at the target organisation. Success in securing access to a dataset could be down to the researcher's ability to demonstrate or convince the target organisation that both user privacy and business privacy will be maintained. In practice (based on the datasets we have observed), the process to ensure privacy is often achieved by limiting or restricting in some way the data that are extracted (either by limiting the features or by specifying more restrictive boundary conditions).

In this context, a social network schema can be useful to assist in understanding and communicating which features are important and which features are not necessary. The pervasiveness of organisational electronic data capture means that these organisations are really useful resources for good social network datasets. Ranking features may lead to a better outcome for the desired research, where key features are not lost during the extraction. For example, two notable ways to reduce the information in a dataset are: (1) remove (through sampling) a number of nodes from the dataset, or (2) remove attribute information about each node. An example of the first approach is described in Blondel et al. (2012); it involved selecting a random sample of 5,000 nodes from the complete mobile phone call network and iteratively adding their first and second degree connections. An example of the second is described in McAuley and Leskovec (2012), which involved no sampling, but completely anonymised the attributes of each of the nodes. Both approaches are completely different ways to address the same problem. Both approaches limit, in different ways, the types of analyses the data can support. If the researcher has a preference to apply a specific type of analysis, then ranking *node attributes* relative to the *sampling technique* may be a useful way to ensure that a source supports the objectives of the research.

Further, if it can be shown that privacy can easily be maintained given a specific extraction (perhaps through a

case study or existing example from a different case/dataset), then we may use a schema to outline how to extract a similar dataset. Going through this process may increase confidence in the method (as it has worked successfully in the past). For example, Brozovsky and Petricek (2007) have published a dataset from the dating website Líbímseti.cz, and describe in detail the node attributes selected (including how the complete set maintains privacy). This dataset was released in 2007, and since then there has (to our knowledge) been no complaint about privacy violation. This fact could form an argument for those seeking a similar dataset from other data sources by demonstrating that the data structure did in fact maintain an acceptable level of privacy.

In both cases, the success in obtaining a rich social network dataset comes down to the ability to: (a) clearly specify and (b) demonstrate privacy features to a third party. Both of these can easily be achieved by proposing a structure using the schema. In addition, given the structure of the schema, it can also be used to facilitate a two-way conversation to agree on a compromise if providing data according to the original request is judged to be too risky.

Finally, it is worth mentioning that the change in size of datasets has presented a new set of challenges. Traditional network statistical approaches are often quite computationally intensive, and as the number of nodes in a network increases, the computational power may increase at an exponential rate. The result of this is that large datasets simple cannot be analysed using traditional approaches. This is becoming an increasingly important consideration when designing a network dataset.

3.3 Assess the suitability of an existing social network dataset

The final route to accessing a social network dataset is through locating and assessing the suitability of an existing dataset for a new purpose. For example, we may want to test a new algorithm or statistical method on a different dataset. In general, there are two possible challenges to achieving this objective: (1) another dataset with the features required does not exist or (2) it is difficult to search for and locate another dataset with the desired features. The schema which we have outlined here allows us to describe, in a standardised way, the features of a dataset. This allows for easy and quick communication through multiple channels (e.g. email, group lists or at conferences) the features that are desired, and hence increase the probability of locating a dataset. In addition, many research labs are beginning to host websites where they describe their publically available network datasets. If these research groups or labs can transition to describing their datasets in a common way (using a schema), then it will greatly help others in locating and identifying rich datasets that suit their requirements. The benefit from using a schema in this context is that it allows for quick and easy communication of features or dataset characteristics. Finally, when documenting or describing data in articles or papers where there is no intention (due to the aforementioned concerns) to release the data to other researchers, a schema can assist in describing fully the features of the dataset in an accessible way. Doing this will allow others to easily understand the structure, assumptions and design decisions in the dataset and hence draw some conclusions about the veracity of the approach without actually viewing the dataset.

4 Conclusion

Today, more than ever, social network analysis approaches to understanding behaviour are popular. This popularity is among those working within traditional social network disciplines like sociology, but also extends more and more to other disciplines, including computer scientists and statisticians. This is in part due to a trend for research to leverage technological advances to create richer datasets, which have many features. This trend has led to a proliferation of rich datasets. Having collected and studied the characteristics of many different social network datasets, we have identified a number of common structural features and characteristics that datasets have. These features include node attributes, data for multiple observation periods (dynamic network data), parallel event data, specific boundary conditions, and known communities. From these features, we propose a social network dataset schema. This schema is designed to support the researcher in documenting, communicating and comparing social network datasets through the use of a standard set of dataset descriptors (describing features). Our proposed schema supports the capture of the underlying informational structure and design decisions of a dataset and assists communication and sharing of social network datasets between researchers (irrespective of their discipline or location).

We discussed how researchers in different contexts could use the schema to validate their data collection strategy or assist them in sourcing or creating richer social network data. These contexts included: validating a proposed social network data collection design and strategy prior to execution; guiding others (who are holding proprietary or sensitive data) on which features are important to help them create an anonymised, aggregated or a sample dataset without losing the most valuable features; or locating a social network dataset with specific features.

This work was based upon an analysis of over 150 social network datasets, prepared by the Dynamics Lab at

University College Dublin. This repository of datasets has been made public, and is available on the Dynamics Lab website at http://dl.ucd.ie.

Our call to action is this: if our objective is to move toward the development of network science as one science, we must adopt a standard way of describing and communicating the complexities (design decisions and assumptions) in assembling a social network dataset. Before publishing, or better yet before collecting data, consider using the approach above to describe the features and characteristics of your dataset. If you are designing a social network dataset, consider capturing the attributes of it as part of the collection process so that when you are ready to share it (or wish to engage others) you will have the tool to easily and quickly achieve this goal.

Acknowledgments This research is funded under Irish Government PRTLI Cycle 5 Simulation Sciences Programme and is co-funded under the European Regional Development Fund of the European Union.

References

Adamic LA, Glance N (2005) The political blogosphere and the 2004 U.S. election. In: Proceedings of the 3rd international workshop on Link discovery—LinkKDD'05. ACM Press, New York, pp 36–43. doi:10.1145/1134271.1134277

Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak C, Ives Z (2007) DBpedia: a nucleus for a web of open data. In: Proceedings of the 6th international 'The semantic web' and 2nd Asian conference on Asian semantic web conference, ISWC'07/ASWC'07. Springer, Berlin, pp 722–735

Batagelj V, Mrvar A (2006) Pajek datasets. http://vlado.fmf.uni-lj.si/pub/networks/data/

Batagelj V, Mrvar A, de Nooy W (2008) Exploratory social network analysis with Pajek. Cambridge University Press, England

Blondel VD, Esch M, Chan C, Clerot F, Deville P, Huens E, Morlot F, Smoreda Z, Ziemlicki C (2012) Data for development: the D4D challenge on mobile phone data. arXiv preprint arXiv:1210.0137

Boyd D, Ellison N (2007) Social network sites: definition, history and scholarship. J Comput Mediat Commun 13(1):210–230. doi:10.1111/j.1083-6101.2007.00393.x

Brozovsky L, Petricek V (2007) Recommender system for online dating service. arXiv preprint cs/0703042

Cross RL, Parker A (2004) The hidden power of social networks: understanding how work really gets done in organizations. Harvard Business School Press, US

De Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, New York

Farrugia M, Hurley N, Payne D, Quigley A (2011) Social network construction in the information age: views and perspectives. In: Ting IH, Hong ZP, Wang LSL (eds) Social network mining, analysis and research trends: techniques and applications. IGI Global, Pennsylvania. doi:10.4018/978-1-61350-513-7

Freeman L, Freeman S (1980) A semi-visible college: structural effects on a social networks group. In: Henderson MM, MacNaughton MJ (eds) Electronic communication: technology and impacts. Westview Press Inc, Boulder, pp 77–85

Giles L, Smith M, Yen J, Zhang H (eds) (2010) Advances in social network mining and analysis, vol 5498. Springer, Berlin. doi:10.1007/978-3-642-14929-0

Gjoka M, Kurant M (2010) Walking in Facebook: a case study of unbiased sampling of OSNs. In: INFOCOM, 2010 Proceedings IEEE. IEEE, pp 1–9

Greene D, Cunningham P (2013) Producing a unified graph representation from multiple social network views. In: Proceedings of ACM Web Science

Hennig M, Brandes U, Pfeffer J, Mergel I (2013) Studying social networks: a guide to empirical research. Campus Verlag GmBH, Frankfurt

Isella L, Stehlé J, Barrat A, Cattuto C, Pinton J-F, Van den Broeck W (2011) What's in a crowd? Analysis of face-to-face behavioral networks. J Theor Biol 271(1):166–180. doi:10.1016/j.jtbi.2010.11.033

Kapferer B (1972) Strategy and transaction in an African factory: African workers and Indian management in a Zambian town. Manchester University Press, London

Kossinets G (2006) Effects of missing data in social networks. Soc Netw 28(3):247–268. doi:10.1016/j.socnet.2005.07.002

Lazega E (2001) The collegial phenomenon: the social mechanisms of co-operation among peers in a corporate law partnership. Oxford University Press, New York

Leskovec J, Huttenlocher D, Kleinberg J (2010) Signed networks in social media. In: Proceedings of the 28th international conference on Human factors in computing systems—CHI'10. ACM Press, New York, p 1361. doi:10.1145/1753326.1753532

Loomis CP, Morales JO, Clifford RA, Leonard OE (1953) Turrialba: social systems and the introduction of change. The Free Press, Glencoe

Maniu S, Abdessalem T, Cautis B (2011) Casting a web of trust over Wikipedia: an interaction-based approach. In: Proceedings of the 20th international conference companion on World wide web, WWW '11. ACM, New York, pp 87–88

McAuley J, Leskovec J (2012) Learning to discover social circles in ego networks. Adv Neural Inf Process Syst 25:548–556

Narayanan A, Shmatikov V (2009) De-anonymizing social networks. In: 2009 30th IEEE symposium on security and privacy. IEEE, pp 173–187. doi:10.1109/SP.2009.22

Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74(3):036104. doi:10.1103/PhysRevE.74.036104

Newman MEJ (2010) Networks: an introduction. Oxford University Press, England. doi:10.1093/acprof:oso/9780199206650.001.0001

Opsahl T, Panzarasa P (2009) Clustering in weighted networks. Soc Netw 31(2):155–163. doi:10.1016/j.socnet.2009.02.002

Palla G, Farkas IJ, Pollner P, Derényi I, Vicsek T (2008) Fundamental statistical features and self-similar properties of tagged networks. New J Phys 10(12):123026. doi:10.1088/1367-2630/10/12/123026

Prosper Marketplace (2010) Prosper data export. http://www.prosper.com/tools/DataExport.aspx. October 2010. v1.2.6

Read K (1954) Cultures of the central highlands, New Guinea. Southwest J Anthropol 10:1–43

Sampson S (1969) Crisis in a cloister. Dissertation, Cornell University

Seierstad C, Opsahl T (2011) For the few not the many? The effect of affirmative action on presence, prominence, and social capital of women directors in Norway. Scand J Manag 27(1):44–54

Stehlé J, Voirin N, Barrat A, Cattuto C, Isella L, Pinton J-F, Vanhems P (2011) High-resolution measurements of face-to-face contact patterns in a primary school. PloS One 6(8):e23176. doi:10.1371/journal.pone.0023176

Van De Bunt GG, Van Duijn MAJ, Snijders TAB (1999) Friendship networks through time: an actor-oriented dynamic statistical

network model. Computat Math Organ Theory 5(2):167–192. doi:10.1023/A:1009683123448

Van Duijn MAJ, Zeggelink EPH, Huisman M, Stokman FN, Wasseur FW (2003) Evolution of sociology freshmen into a friendship network. J Math Sociol 27(2–3):153–191. doi:10.1080/00222500305889

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Structural analysis in the social sciences, vol 8. Cambridge University Press, London

Zhao B, Sen P, Getoor L (2006) Entity and relationship labelling in affiliation networks. In: ICML workshop on Statistical Network Analysis

Zheleva E (2011) Prediction, evolution and privacy in social and affiliation networks. PhD Dissertation, University of Maryland, College Park