Data Descriptor

LOSD: A Linked Open Social Dataset for Scientific Benchmarking

Renato Fabbri [1]* and Osvaldo N. Oliveira Junior [2]

[1] Institute of Mathematical and Computer Sciences, University of São Paulo (ICMC/USP), São Carlos, Brazil; renato.fabbri@gmail.com
[2] São Carlos Institute of Physics, University of São Paulo (IFSC/USP), São Carlos, Brazil; chu@ifsc.usp.br
* Correspondence: renato.fabbri@gmail.com

Version July 1, 2019 submitted to Data

Abstract: The fields of social network analysis and complex networks are widely researched. In fact, a myriad of results have been reported which are based in diverse data, most often not accessible to researchers other than the publishing authors. In order to provide the scientific community a common repertoire, this work presents an open dataset with diverse provenance of human social online networking. Current data was obtained in 2013-16 from well-known online social networking platforms: Facebook, Twitter, IRC, Email. Data from ParticipaBR, AA, and Cidade Democrática are also supplied as samples of less known instances. The linked data representation was adopted in order to comply with current best practices, homogenize access, facilitate discovery and analyzes which integrate third party and provided data. Futhermore, the method for obtaing and making the data available is potentially novel, and may be of use for other research groups. This document presents a description, usage remarks and overall statistics of the dataset, which should favor subsequent work in complex and social networks and enhancements to the dataset.

Dataset: https://rfabbri.linked.data.world/d/linked-open-social-data/

Keywords: social networks; complex networks; linked data; semantic web; big data; benchmark data; Facebook; Twitter; irc; email

1. Summary

The research on human social and complex networks often rely on data that express real phenomena [1–3]. Such data is regularly not made publicly available, which hinders validation of the results and does not favor subsequent work. Furthermore, there is a lack of open datasets for benchmarking results related to complex and social networks, yielding diverse results from poorly related sources. Social networks linked datasets are currently most often restricted to a specific platform or express statistics of social networks, as described in Section 1.1. This work presents a dataset named LOSD (Linked Open Social Data) that has been achieved by linking social data derived from diverse provenance, namely: Facebook, Twitter, IRC, Email groups, ParticipaBR, Cidade Democrática, and AA. Such data may be of use as a common and public repertoire for scientific research involving complex networks, specially if social networks and textual content are pinpointed. In fact, the data has already been used in scientific research [4,5] (CNPq grant 140860/2013-4), and should be further used e.g. in researching complex networks and data visualization (FAPESP grant 2017/05838-3) which relies on networks enriched by textual data and on dynamic (or time-evolving) networks. Data collection was performed by a potentially innovative method which places the researcher at the core of data collection, and keeps the resources (e.g. code, data, documents) open, in order to ammeliorate ethical issues involved in studying human social systems.

This paper is organized as follows. Next subsection briefly discusses related work. Section 2 describes the data currently available. Section 3 describes the methods used for data gathering,

translation into RDF linked data, and the synthesis of auxiliary structures. Section 4 holds directions for using the data.

1.1. Related work

Most available datasets on human social networks rely only on one source (e.g. [6,7]), on specialized communities such as related to research or movies (e.g. [8], or present only statistics about social networks, and not the data that yields the networks (e.g. [9]. Linked human social network data is scarcely found in the scientific community. In [10], the authors presented a set of tools to obtain linked social data available in the public data cloud, and made available a dataset from the Advogato platform, dedicated to free software developers linked by trust relatioships. There are various datasets which makes available data obtained from Twitter (e.g. [11,12]), most often with a specific focus (users, retweets, hashtags, etc) and not presented to the scientific community by means of a research article. The datasets which are most closely related to LOSD are KONECT [13], SNAP Datasets [14], the Network Repository [15], and ICON [16]. None of them are (RDF) linked data but express the networks in formats specialized for networks. In total, these datasets make available thousands of social networks, from diverse provenance and of various types (e.g. simple, directed, weighted, bipartite, temporal networks). Compared to LOSD, KONECT, SNAP, Network Repository, and ICON are complementary, providing a further diversity of networks, but only as simplified data, i.e. in specialized formats, most often expressed as columns of node ids that are connected and metadata (such as timestaps for temporal networks). In summary, no linked human social network dataset presented to the scientific community containts the diverse provenance used in LOSD, or provides such a wide range of metadata, potentially because researchers did not have the means to acquire and share the data as described in Section 3.1, and because linked data standards are (still) not widely adopted in every research domain.

2. Data Description

2.1. Data outline

The database consists of 31,996,740 triples, 3,030,433 network links yield by main interactions or relations, 350,284 participants and 245,194,377 characters. Among all snapshots, 63 are ego snapshots, 54 are group snapshots; 50 have interaction edges, 89 have friendship edges; 43 have text content from messages. Section 4.3 holds the queries used to obtain such data. This section describes de data in LOSD in very brief terms. The Supplementary document is provided for further information about the dataset.
Data was gathered from:

- public APIs (Twitter, Email);
- public logs (IRC and AA);
- Netvizz software [17] and subsequent donation by users (Facebook);
- donation by system administrators (AA, ParticipaBR, Cidade Democrática).

2.2. Snapshots

Of central importance to the presented dataset is the concept of a snapshot. A snapshot is herein a set of data gathered together, at a contiguous time unit. Examples: the first 20 thousand email messages of an email list comprises a snapshot; the tweets from the MAMA music event is a snapshot; the friendship, interaction and posts structures of a facebook group, prospected at the same time, is a snapshot. Table 1 holds the number of snapshots from each provenance in LOSD.

2.3. Facebook data

Friendship ego networks (networks whose nodes represent friends of a user and links represent friendships) were donated from individual users in 2013 and 2014. Friendship and interaction networks from groups were gathered from groups where the first author was a participant. Additionally, some groups have post texts along some metadata, such as the number of likes.

Table 1. Number of snapshots from each provenance.

social protocol	number of snapshots
Algorithmic Autoregulation	3
Cidade Democrática	1
Email	4
Facebook	88
IRC	4
ParticipaBR	1
Twitter	16
all	117

2.4. Twitter data

Tweets were gathered through the Twitter streaming public API. Each snapshot is unified by a distinct hashtag. Network links are canonically yield by retweets, but replies and user mentions are also kept in the database. LOSD kept the main data fields returned by Twitter API, such as creation date, text, and hashtags.

2.5. IRC data

Public IRC (Internet Relay Chat) logs were used to obtain IRC snapshots. LOSD holds records of relations yield by directed messages and mentions and the text related to each message.

2.6. Email data

Email snapshots refer to individual email lists. All messages were obtained from the Gmane public email archive [18]. Each message has the original text and the text without some of the lines from previous messages or that are software code or author signature. Most importantly, each message instance holds the ID of the message it is a reply to, if any, which enables the achievement of canonical interaction networks [4,19].

2.7. ParticipaBR data

ParticipaBR is a Brazilian federal platform for social participation, the most important platform from the myriad of social participation platforms that were available in 2011-2018 [20]. Texts are derived from blog posts and networks may be obtained from both friendship and interaction criteria.

2.8. Cidade Democrática data

Cidade Democrática is a Brazilian civil society social participation portal. Data gathered is complex in the number of types of instances as noticeable in the Supplementary document.

2.9. AA data

AA (Algorithmic Autoregulation [21,22]) is a software to keep track of dedications by sharing sentences about ongoing work and testifying about their quality. The data was gathered from different versions of the system and from an IRC log.

3. Methods

LOSD is expressed as linked data through RDF and is ontologicaly described through the data-driven ontology synthesis method presented in Section 3.5.

3.1. Data acquisition

Research involving human subjects raise well-known ethical issues, although ethical regulation for (non-medical) research in social systems is not always endorsed by scientists [23]. Furthermore, the online social networking platforms often restrict access to the data and user may consider illegal the use of their data through web crowlers [24–26]. Although Twitter and GMANE (email) data is available through public APIs, the other sources are more problematic. In this context, research using data, e.g.

from Facebook, is precarious if not prohibited. We reached a potential solution to this issue through continuated research with collaborators in the fields of phylosophy, social sciences, antropology and psychology. The conceptual framework is presented in [27,28], and in essence consists in researching first the social networks related to the researchers, and keeping the procedures, data, algorithms and results publicly available. These practices enabled collecting and sharing the data in LOSD, and was called "Anthropological Physics" because it relies in the etnographic tecnique of investigating the social structures by studying the researcher, as to ammeliorate the ethical issues (not to be confused e.g. with "virtual ethnography" [29]). Accordingly, aside from the data acquired from Twitter, all data in LOSD was gathered from social instances where the first author is a participant. For example, all Facebook ego networks were donated to the first author by users which know him and were interested in his research and intervention on the networks, and all email and Facebook groups data was obtained from groups where the first author is a member.

Facebook data was obtained through the Netvizz software [30], Facebook allowed downloading friendships and interaction data as available in LOSD. Data from Twitter was obtained through the Twitter streamming API. Data From IRC was obtained through bots that kept logs of the messages. Email data was obtained through the GMANE public API. Data from ParticipaBR, Cidade Democrática, and AA was obtained through contact of the researcher with the developers of the platforms.

3.2. Linked open data

Linked data refers to data published in the web in such a way that it is machine readable and conforms to a set of best practices. The yielded web of data is composed of documents on the web such as the web of HTML documents. In practice, the publication of linked data can me summarized by 1) the use of RDF and 2) the use of RDF links to interlink data from different sources [31,32]. The web is expected to be interconnected and to grow by the systematic application of four steps [33]:

- Use URIs to identify things [34].
- Use HTTP URIs.
- Provide useful information when a URI is accessed via HTTP.
- Provide other URIs in the description of resources so human and machine agents can perform discovery.

The Linked Open Data [35] cloud is a constantly growing cloud of data, the global data space, which is usually conceived as centered around the DBPedia, a linked data representation of data from Wikipedia [36,37], and sometimes called the Giant Global Graph (GGG, akin to the WWW).

3.3. RDF

The Resource Description Framework (RDF), a W3C recommendation, is a model for data interchange. It is based on the idea of making statements about resources in the form of triples, i.e. expressions in the form "subject - predicate - object". RDF can be serialized in several file formats, including RDF/XML, Turtle and Manchester, which all, in essence, represent a labeled and directed multi-graph. RDF may be stored in a type of database called a triplestore. [38]

As an example of an RDF statement, the following triple in the Turtle format asserts that "the paper has color white":

```
http://example.org/Thing#Paper http://example.org/hasColor
http://example.org/Color#White .
```

3.4. RDFS and OWL ontologies

One of the most important notions of the semantic web is that of an ontology [39]. An ontology, in this context, is a formalized conceptualization, comprised by concepts and relations between the concepts and between the relations themselves. In current semantic web, most simple ontologies are written using the RDFS protocol, by which one can specify, among other things [40]:

- concepts;
- properties, which are concepts that are used as predicates and thus relate concepts;
- special relations between concepts that state that one concept is more general than the other[1];

[1] This kind of relation is called hypernymy. Examples: mammal is a hypernym of monkey, drink is a hypernym of beer.

- the subjects and objects that can occurs in a triple where a specific property is a predicate.

One can also write an ontology using the OWL protocol, with which all the expressive capabilities of RDFS are available, but one can also, among other things [39]:

- state "property axioms", i.e. specify if a property is e.g. reflexive or transitive;
- state "class restrictions", i.e. specify if a class instance e.g. necessarily holds a relation to another class instance or data.

OWL has a richer vocabulary than RDFS and is (way more) complex. This complexity is a drawback together with the greater computational cost for performing inference. The advantage is the greater power to represent conceptualizations. Using ontologies enables automated reasoning. For example, if a property `:r` is known to relate only mokeys to trees, then, if there is a triple `:a :r :b`, the instance `:a` is considered a monkey, and instance `:b` is considered a tree, even if there is no explicit declaration similar to `:a rdf:type :Mokey`.

For the data described in this article, the following features are most relevant [39,40]:

- Using RDFS, one may define a property domain, i.e. the class instances that are allowed as the object in a triple with the property.
- Using RDFS, one may define a property range, i.e. the class instances that are allowed as the predicate in a triple with the property.
- Using OWL, one may define an existential restriction for a class `:C`, i.e. declare that an instance `:C#I` of such class has at least one relation with a property `:p` and a class `:C2` in the form `:C#I :p :C2#I`, where `:C2#I` is an instance of the class `:C2`.
- Using OWL, one may define an universal restriction for a class `:C`, i.e. declare that any relation of an instance `:C#I` of such class, with a property `:p` of the form `:C#I :p :CX#I`, implies that `:CX#I` is an instance of a specific class, e.g. `:C2`.

3.5. Data-driven ontology synthesis

Most ontologies are created by domain experts [39] even though the data they arrange is often given by a software system and has thus a predefined structure. We developed a simple ontology synthesis method that probes the ontological structure in the data with SPARQL queries and post-processing. The results are OWL code and diagrams which are available in the Supplementary document of this article. The method can be extended to comprise further OWL axioms and restrictions, but is currently performed to fit present needs with maximum simplicity. Present needs are limited to informative figures and the steps implemented are as follows:

1. Obtain all distinct classes with the query:

```
SELECT DISTINCT ?class_uri WHERE { ?s a ?class_uri .}
```

2. For each class, obtain the properties that occur as predicates in triples where the subject is an instance of the class:

```
SELECT DISTINCT ?property_uri WHERE { ?s a <class_uri> . ?s ?property_uri ?o .}
```

3. Compare the total number of individuals (`?cs1`) of the class (`class_uri`) with the number of such individuals (`?cs2`) that are subjects of at least one triple where the predicate is the property (`property_uri`) and the object is an instance of the same class (`class2_uri`). The queries are:

```
SELECT (COUNT(DISTINCT ?s) as ?cs1) WHERE { ?s a <class_uri> }
SELECT DISTINCT ?oc WHERE {
  ?s a <class_uri> . ?s <property_uri> ?o . ?o a ?oc .
}
```

Then, for each object class (`object_curi`) in `?oc`:

```
SELECT (COUNT(DISTINCT ?s) as ?cs2) WHERE {
  ?s a <class_uri> . ?s <property_uri> ?o . ?o a <object_curi> .
}
```

When `?cs1` matches `?cs2`, an existential restriction is found. And with the following query, when `?cs` is 0, a universal restriction is found:

```
SELECT (COUNT(DISTINCT ?s) as ?cs) WHERE {
  ?s a <class_uri> . ?s <property_uri> ?o . ?o a ?oc .
  FILTER(str(?oc) != 'object_curi')
}
```

4. Retrieve all object classes or datatypes where the subject is an instance of the class and the predicate is the property:

```
SELECT DISTINCT ?co (datatype(?o) as ?do) WHERE {
  ?s a <class_uri>. ?s <property_uri> ?o . OPTIONAL { ?o a ?co . }
}
```

5. Obtain all distinct properties:

```
SELECT DISTINCT ?p WHERE { ?s ?p ?o }
```

6. For each property, check if it is functional, i.e. if it occurs at most once with each subject. This is performed by counting the objects and further verifying that they are at most one. The query is:

```
SELECT DISTINCT (COUNT(?o) as ?co) WHERE {
  ?s <property_uri> ?o
} GROUP BY ?s
```

7. For each property, find the incident range (possible subjects) and domain (possible objects) with the queries:

```
SELECT DISTINCT ?co (datatype(?o) as ?do) WHERE {
  ?s <property_uri> ?o . OPTIONAL { ?o a ?co . }
}
SELECT DISTINCT ?cs WHERE { ?s <property_uri> ?o . ?s a ?cs . }
```

8. Render diagrams to achieve figures as exemplified in Section 4.2 and in the Supporting Information document.

3.6. Standardization

LOSD data is represented through URIs and triples by translation of various formats to RDF. URIs are built in the namespace http://purl.org/socialparticipation/participationontology/ which are identified herein with the prefix "`po:`". Classes and properties are built by adding a suffix to the root, as in `po:Participant` or `po:text`. Classes have "UpperCamelCase" suffixes while properties have "lowerCamelCase" suffixes. All class instances, such as participants, messages, friendships and interactions, are linked to snapshots through the triple `<instance> po:snapshot <snapshot_uri>`. Message texts, including comments, are objects in a triple: `<message_id> po:text '...the text...'`. Preprocessed texts are objects of triples: `<message_id> po:cleanText '...te text...'`. More specialized predicates are used for text storage when necessary, such as `po:htmlBodyText` and `po:cleanBodyText` used for ParticipaBR articles. A participant URI is unique throughout the provenance (e.g. the same for the same participant in all Twitter snapshots). To enable annotations which differ when the snapshot changes, `po:Observation` class instances are used in the triple `<participant_uri> po:observation <observation_uri>`. Observation instances are then linked to the snapshot and the data.

Instance URIs are built using the URI from the class they derive from plus a hashtag character, a provenance string (e.g. `facebook-legacy` or `participabr-legacy`), and an identifier; e.g. `po:Participant#<provenance-legacy>-<id>`. All snapshot URIs follow the formation rule: `po:<SnapshotProvenance>#<snapshot_id>`. All snapshot ids follow the formation rule: `<platform>-legacy-<further_identifier>`; e.g. `irc-legacy-labmacambira` or `email-legacy-linux.audio.devel1-20000`.

The authors chose to express all URIs in LOSD locally (i.e. within a local namespace), and linkage with external namespaces may be achieved e.g. by using subclass and "same as" relations. Also, each snapshot holds specific `po:observation` for each participant, allowing for specific metadata such as provided by the original data or as may be provided uppon analysis. LOSD does not hold data about each snapshot (such as statistics of the networks), and linkage to external namespaces (such as FOAF and Dublin Core) to avoid overcomplicating this initial contribution and unecessary use of storage space.

4. Usage notes

4.1. Triplification routines

For each social platform there is a *triplification* routine, i.e. a script for expressing the data as RDF. Original formats and further information are presented in Table 2.

Table 2. Social platforms, original formats and further observations for the database.

social platform	original format	further observations	toolbox
AA	MySQL and MongoDB databases; IRC text logs	donated by AA users	Participation [41]
Cidade Democrática	MySQL database	donated by admins	Participation
Email	mbox	obtained through the Gmane public archive	Gmane [18]
Facebook	GDF, GML and TAB	obtained through Netvizz [17]	Social [42]
IRC	plain text log	obtained through Supybot logging	Social
ParticipaBR	PostgreSQL database	donated by admins	Participation
Twitter	JSON	obtained through Twitter streaming API	Social

4.2. Diagrams of the data and auxiliary tables

LOSD navigation may be assisted by diagrams which describe the structure from each provenance. Such diagrams are in the Supporting Information document together with tables to facilitate the appreciation of the provided data. A very simple example is given in Figure 1 where the friendship structure of the Facebook snapshots is revealed.

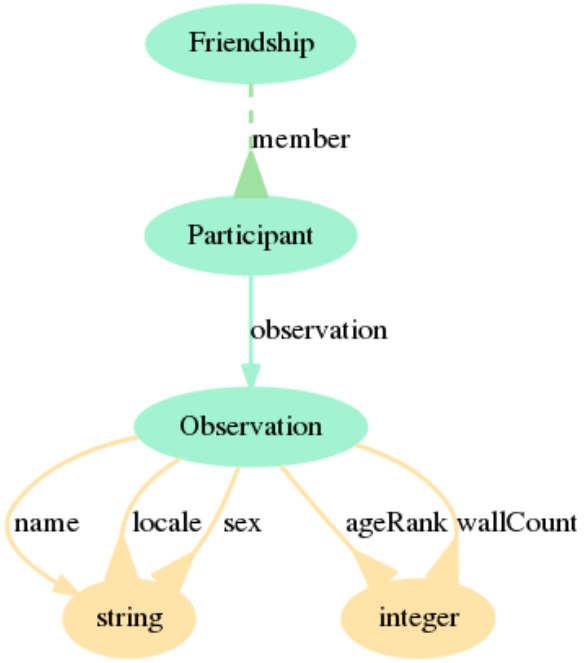

Figure 1. A diagram of the core structure that yields the friendship networks of the Facebook snapshots. A green edge denotes an OWL existential class restriction; an inverted nip denotes an OWL universal class restriction; a full (non-dashed) edge denotes an OWL functional property axiom (see Section 3.4). Further information and complete diagrams for each provenance are available in the Supporting Information document.

4.3. SPARQL queries

There are numerous useful and general purpose SPARQL queries to be performed against LOSD. In this section, a selection of some of the most basic queries are made available due to their potential to be varied. All queries assume the preamble `PREFIX po: <http://purl.org/socialparticipation/po/>`.

- Retrieve the number of participants:

```
SELECT (COUNT(DISTINCT ?author) as ?c) WHERE {
  ?author a po:Participant .
}
```

- Retrieve the number of relations, be them interactions or friendships:

```
SELECT (COUNT(?interaction) as ?c) WHERE {
  { ?interaction a po:Friendship } UNION { ?interaction a po:Interaction }
  UNION { ?interaction po:retweetOf ?message }
  UNION { ?interaction po:replyTo ?message }
  UNION { ?interaction po:directedTo ?participant }
}
```

- Retrieve all text produced by a specific user with URI `user_uri`:

```
SELECT (CONCAT(?text) as ?texts) WHERE {
  ?activity po:author <user_uri> . ?activity po:text ?text .
}
```

- List 1000 users (URIs and names) with the most friendships and the number of friendships in descending order by the number of friendships:

```
SELECT DISTINCT ?participant (COUNT(?friendship) as ?c) WHERE {
  ?friendship a po:Friendship . ?friendship po:member ?participant .
} ORDER BY DESC(?c) LIMIT 1000
```

- Retrieve text messages with the word “crucial” (case insensitive):

```
SELECT ?text WHERE {
  ?activity po:text ?text . FILTER regex(?text, 'crucial', 'i')
}
```

- List participants and respective full names whose name has the substring "Amanda":

```
SELECT DISTINCT ?participant ?name WHERE {
  ?participant po:observation ?obs . ?obs po:name ?name .
  FILTER regex(?name, 'Amanda', 'i')
}
```

- Return all pairs of friends of a participant (with URI `participant_uri`) which are friends themselves:

```
SELECT DISTINCT ?friend1 ?friend2 WHERE {
  ?friendship1 po:member <participant_uri> . ?friendship1 po:member ?friend1 .
  ?friendship2 po:member <participant_uri> . ?friendship2 po:member ?friend2 .
  ?friendship3 po:member ?friend1 . ?friendship3 po:member ?friend2 .
}
```

- Return all interactions from replies in a snapshot:

```
SELECT ?from ?to WHERE {
  ?message1 po:snapshot <snapshot_uri> .  ?message2 po:replyTo ?message1 .
  ?message1 po:author ?from .  ?message2 po:author ?to .
}
```

4.3.1. Obtaining the networks

A complex network $N = (n, l, m)$ may be regarded essentially as a set of nodes n, a set of links l that related the nodes, and metadata m such as node and link weights, text related to the nodes, etc. This section presents queries for obtaining networks using LOSD which are promptly envisioned by the authors. Many other networks may be obtained because the criteria that entail links, and restrains in the set of nodes and of links, are determined by the research or application being developed. Furthermore, we do not focus here on multilayer networks (e.g. bipartite networks), or on multigraphs (or muldimensional networks), i.e. on networks with more than one type of node or of link. Here is a list with short descriptions of the networks and the corresponding SPARQL queries:

- Overall facebook friendship network available in LOSD:

```
SELECT ?a1 ?a2 WHERE {
  ?f a po:Friendship .  ?f po:member ?a1 . ?f po:member ?a2 .
}
```

- Overall facebook friendship network derived only from ego networks:

```
SELECT ?a1 ?a2 WHERE {
  ?f a po:Friendship . ?f po:snapshot ?s . ?s po:isEgo True .
  ?f po:member ?a1 . ?f po:member ?a2 .
}
```

- Overall facebook interaction network available in LOSD:

```
SELECT ?a1 ?a2 WHERE {
  ?f a po:Interaction .
  ?f po:interactionFrom ?a1 . ?f po:interactionTo ?a2 .
}
```

- Email interaction network for email list shapshot with URI `http://xxxx`, with the all (preliminarily cleaned) text related to each message (node atribute):

```
SELECT ?a1 ?a2 ?t1 ?t2 WHERE {
  ?m po:snapshot <http://xxxx> . ?m a po:EmailMessage .
  ?m po:author ?a1 . ?m po:replyTo ?m2 . ?m2 po:author ?a2 .
  ?m po:cleanText ?t1 . ?m2 po:cleanText ?t2 .
}
```

- IRC users related by messages directed to each other, and the text of each message (link attribute):

```
SELECT ?a1 ?a2 ?t WHERE {
  ?m a po:IRCMessage .  ?m po:author ?a1 . ?m po:directedTo ?a2 .
  ?m po:cleanText ?t .
}
```

- Twitter users related by retweets:

```
SELECT ?a1 ?a2 WHERE {
  ?m1 po:retweetOf ?m2 . ?m1 po:author ?a1 . ?m2 po:author ?a2 .
}
```

- Twitter users related by user mentions and the text of the tweets:

```
SELECT ?a1 ?a2 ?t WHERE {
  ?m a po:Tweet . ?m po:author ?a1 . ?m po:userMention ?a2 .
  ?m po:text ?t .
}
```

- Friendship network from ParticipaBR, with the date when each frienship was established:

```
SELECT ?a1 ?a2 ?d WHERE {
  ?s po:shapshotID 'participabr-legacy' . ?f po:snapshot ?s .
  ?f a po:Friendship .
  ?f po:member ?a1 . ?f po:member ?a2 . ?f po:createdAt ?d .
}
```

- Interaction network from ParticipaBR obtained by comments on articles:

```
SELECT ?a1 ?a2 WHERE {
  ?s po:shapshotID 'participabr-legacy' . ?a po:snapshot ?s .
  ?a a po:Article .
  ?a po:author ?a1 . ?c po:article ?a . ?c po:author ?a2 .
}
```

- Interaction network from ParticipaBR obtained by comments and votes on comments:

```
SELECT ?a1 ?a2 WHERE {
  ?s po:shapshotID 'participabr-legacy' . ?c po:snapshot ?s .
  ?c a po:Comment . ?c po:author ?a1 .
  { ?v a po:Vote . ?v po:reference ?c . ?v po:author ?a2 .
  } UNION {
    ?c2 a po:Comment . ?c2 po:replyTo ?c . ?c2 po:author ?a2 .
  }
}
```

- Interaction network from Cidade Democrática obtained by comments on Topics:

```
SELECT ?a1 ?a2 WHERE {
  ?s po:shapshotID 'cidadedemocratica-legacy' . ?t po:snapshot ?s .
  ?t a po:Topic . ?t po:author ?a1 .
  ?c a po:Comment . ?c po:topic ?t . ?c po:author ?a2 .
}
```

- AA users, related by AA session validations:

```
SELECT ?a1 ?a2 WHERE {
  ?s po:author ?a1 . ?s po:checkParticipant ?a2 .
}
```

This is not an exaustive list. For example, other instances which are related in LOSD, and have authors, also yield valid criteria for interaciton networks. For a less canonical example for obtaining networks, links may be derived from sufficiently similar vocabulary used by participants, or from activity with sufficiently similar dates, which yield networks which are not interaction or friendship networks. Figure 2 describes most often criteria used to obtain social networks.

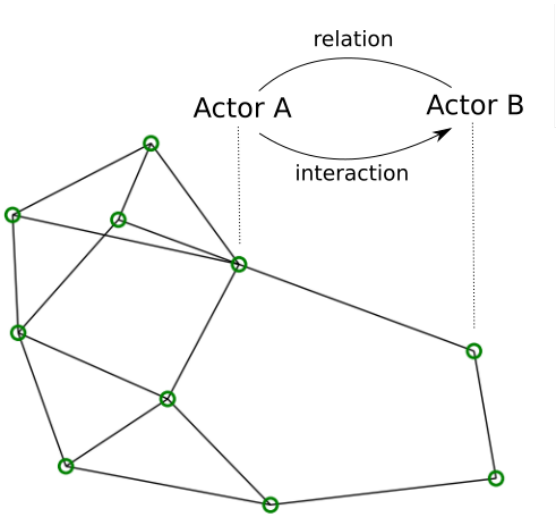

Figure 2. Most usual paradigm for considering social networks: actors (i.e. participants, authors) are linked through interactions (e.g. replies, likes, comments) or relations (e.g. friendships, similar activity patterns).

4.4. Data access

We highly encourage interested parties to take advantage of the API at https://apidocs.data.world/toolkit/api. The user responsible for LOSD has id "rfabbri" and the dataset id is "linked-open-social-data". For making use of the API, and exploiting further services provided by the platform (such as integrations with data analysis software), the researcher needs to register in Data.World, which is fast and free, and facilitates assistance by the platform maintainers. Anonynous (not authenticated) SPARQL queries may be performed using the endpoint at: https://rfabbri.linked.data.world/sparql/linked-open-social-data, kindly provided by the Data.World staff.

4.5. Example network

Figure 3 exemplifies a network obtained through a SPARQL query made to LOSD. The nodes are colored and sized according to network metrics, as explained in the caption. With `http://purl.org/socialparticipation/po/Snapshot#facebook-legacy-CalebLuporini13042013` as the `snaphot_uri`, the query performed to find the network is:

```
PREFIX po: <http://purl.org/socialparticipation/po/>
SELECT ?a1 ?a2 WHERE {
  ?f a po:Friendship . ?f po:snapshot <snapshot_uri> .
  ?f po:member ?a1, ?a2 .
  FILTER(?a1 != ?a2)
}
```

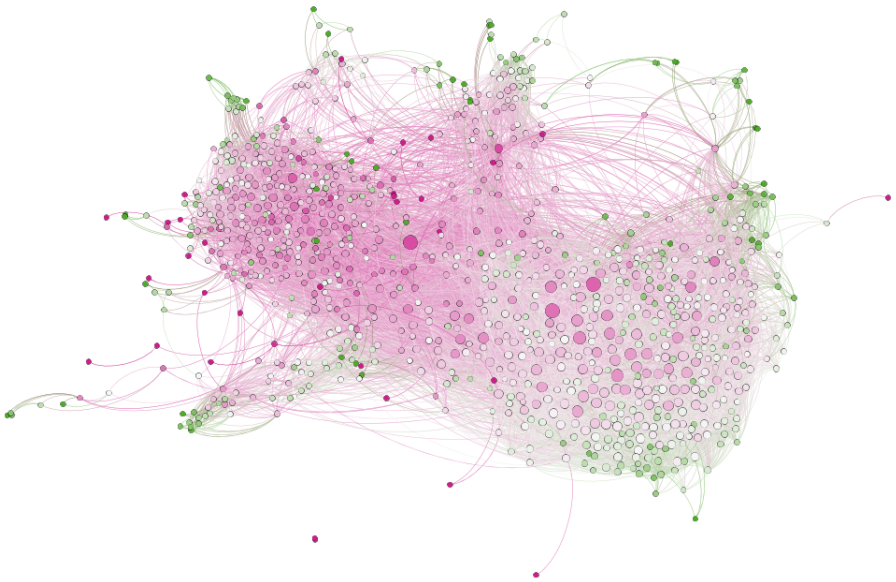

Figure 3. A network retrieved from LOSD as described in Section 4.5. Nodes are colored according to clustering coefficient (greatest values are green, smallest values are pink) and their sizes are proportional to degree (number of links). The network has 1032 nodes and 24653 links, average degree of 47.77, average clustering coefficient of 0.47, diameter of 7 (greatest shortest path between nodes), etc.

5. Conclusions and further work

This article outlined LOSD, a large human online social networking dataset, with almost 32 million triples and diverse provenance. Tools and procedures used to achieve LOSD, and fundamental queries needed to retrieve the data, were also described to the convenience of the newcommer and for a consistent exposition. All data is publicly available, and the scientific community now has consistent and curated data that enables research validation, continuity, and benchmarking. The diagrams and tables of the Supplementary document of this article has further directions on the available structures and is an overview complement.

The authors plan on using LOSD for analyzing simple, bipartite and dynamic (i.e. time-evolving) networks. The research and development of data visualization interfaces using LOSD is envisioned by

the authors (FAPESP grant). Directly related to publications, the authors should submite a short paper describing and exploring the data collection procedure outlined in Section 3.1 and the data-driven ontology synthesis outlined in Section 3.5. For enhancing LOSD, it should be expanded upon need or request from interested parties. Also, software packages for LOSD data retrieval and analysis should be made available. Routines for anonymization of LOSD data may be desired in order to preserve privacy of the participants [43]. Careful linkage of LOSD classes and entities with third party data (e.g. DBPedia or other Data.World datasets) should make LOSD a five start dataset [44].

Supplementary Materials: The following are available online at http://www.mdpi.com/2306-5729/xx/1/5/s1, Figure S1: title, Table S1: title, Video S1: title.

Author Contributions: Conceptualization, writing, formal analysis, R.F; senior supervision, final writing, O.N.O.Jr.

Funding: This research was funded by CNPq (grant 14068/2013-4) and FAPESP (grant 2017/05838-3).

Acknowledgments: The authors acknowledge the financial support of the São Paulo State Research Foundation (FAPESP grant 2017/05838-3) and the National Council for Scientific and Technological Development (CNPq, grant 140860/2013-4). The authors thank Data.World staff for providing the technical assistance when for providing LOSD, for providing the authors the possibility of uploading LOSD which is above their standard size limits, and for providing the authors and the scientific community the public query endpoint mentioned in Section 4.4. The views expressed do not reflect the official policy or position of FAPESP, CNPq or Data.World. The authors thank all participants in the experiments which resulted in the donation of the data they gathered as described in Section 3.1.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Estrada, E. *The structure of complex networks: theory and applications*; Oxford University Press, 2012.
2. Newman, M. *Networks: an introduction*; Oxford university press, 2010.
3. Barabási, A.L. Linked: The new science of networks, 2003.
4. Fabbri, R.; Fabbri, R.; Antunes, D.C.; Pisani, M.M.; de Oliveira Junior, O.N. Temporal stability in human interaction networks. *Physica A: Statistical Mechanics and its Applications* **2017**, *486*, 92–105.
5. Fabbri, R. Topological stability and textual differentiation in human interaction networks: statistical analysis, visualization and linked data. PhD dissertation, Universidade de São Paulo, 2017.
6. Jankowski, J.; Michalski, R.; Bródka, P. A multilayer network dataset of interaction and influence spreading in a virtual world. *Scientific data* **2017**, *4*, 170144.
7. Lewis, K.; Kaufman, J.; Gonzalez, M.; Wimmer, A.; Christakis, N. Tastes, ties, and time: A new social network dataset using Facebook. com. *Social networks* **2008**, *30*, 330–342.
8. Kaminski, J.; Schober, M.; Albaladejo, R.; Zastupailo, O.; Hidalgo, C. Moviegalaxies - Social Networks in Movies, 2018. doi:10.7910/DVN/T4HBA3.
9. Social Media Channels and Statistics at the National Archive, 2014.
10. Petrović, G.; Fujita, H. Soner: Social network ranker. *Neurocomputing* **2016**, *202*, 104–107.
11. Kratzke, N. The #BTW17 Twitter Dataset - Recorded Tweets of the Federal Election Campaigns of 2017 for the 19th German Bundestag, 2017. Funded via general support for research by Lübeck University of Applied Sciences., doi:10.5281/zenodo.835735.
12. Wang, B.; Tsakalidis, A.; Liakata, M.; Zubiaga, A.; Procter, R.; Jensen, E. SMILE Twitter Emotion dataset, 2016. doi:10.6084/m9.figshare.3187909.v2.
13. Kunegis, J. Handbook of Network Analysis [KONECT–the Koblenz Network Collection]. *arXiv preprint arXiv:1402.5500* **2014**.
14. Leskovec, J.; Krevl, A. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data, 2014.
15. Rossi, R.A.; Ahmed, N.K. The Network Data Repository with Interactive Graph Analytics and Visualization. AAAI, 2015.
16. Clauset, A.; Tucker, E.; Sainz, M. The colorado index of complex networks, 2016.
17. Rieder, B. Studying Facebook via data extraction: the Netvizz application. Proceedings of the 5th Annual ACM Web Science Conference. ACM, 2013, pp. 346–355.

18. Fabbri, R. gmane toolbox. https://github.com/ttm/gmane, 2015.
19. Bird, C.; Gourley, A.; Devanbu, P.; Gertz, M.; Swaminathan, A. Mining email social networks. Proceedings of the 2006 international workshop on Mining software repositories. ACM, 2006, pp. 137–143.
20. Peixoto, A.d.C. Instrumentos da democracia participativa: um estudo sobre o Participa. br e o Dialoga Brasil **2016**. Monografia (Bacharel em Gestão de Políticas Públicas), UnB (Universidade de Brasília), Brasília, Brazil.
21. Fabbri, R.; Fabbri, R.; Vieira, V.; Penalva, D.; Shiga, D.; Mendonça, M.; Negrão, A.; Zambianchi, L.; Thumé, G.S. THE ALGORITHMIC AUTOREGULATION SOFTWARE DEVELOPMENT METHODOLOGY/A METODOLOGIA DE DESENVOLVIMENTO DE SOFTWARE AUTORREGULAÇÃO ALGORÍTMICA. *Revista Electronica de Sistemas de Informaçao* **2014**, *13*, 1.
22. Fabbri, R. The Algorithmic-Autoregulation (AA) Methodology and Software: a collective focus on self-transparency. *arXiv preprint arXiv:1711.04612* **2017**.
23. Dingwall, R. The ethical case against ethical regulation in humanities and social science research. *Twenty-First Century Society* **2008**, *3*, 1–12.
24. Determann, L. Social media privacy: a dozen myths and facts **2012**.
25. Al-khateeb, S.; Agarwal, N. *Deviance in Social Media and Social Cyber Forensics: Uncovering Hidden Relations Using Open Source Information (OSINF)*; Springer, 2019.
26. Bechmann, A.; Vahlstrup, P.B. Studying Facebook and Instagram data: The digital footprints software. *First Monday* **2015**, *20*.
27. FABBRI, R. What are you and I? [Anthropological physics fundamentals] **2015**.
28. Antunes, D.; Fabbri, R.; Pisani, M.M. Anthropological physics and social psychology in the critical research of networks. *International Conference on Complex Systems (2015)* **2015**.
29. Hine, C. Virtual ethnography: Modes, varieties, affordances. *The SAGE handbook of online research methods* **2008**, pp. 257–270.
30. Rieder, B. Studying Facebook via data extraction: the Netvizz application. Proceedings of the 5th annual ACM web science conference. ACM, 2013, pp. 346–355.
31. Heath, T.; Bizer, C. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology* **2011**, *1*, 1–136.
32. Polleres, A.; Kamdar, M.R.; Fernandez Garcia, J.D.; Tudorache, T.; Musen, M.A. A more decentralized vision for linked data **2018**.
33. Berners-Lee, T. Design issues: Linked data, 2006.
34. Masinter, L.; Berners-Lee, T.; Fielding, R.T. Uniform resource identifier (URI): Generic syntax **2005**.
35. Umbrich, J.; Decker, S.; Hausenblas, M.; Polleres, A.; Hogan, A. Towards dataset dynamics: Change frequency of linked open data sources **2010**.
36. Bizer, C.; Lehmann, J.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; Hellmann, S. DBpedia-A crystallization point for the Web of Data. *Web Semantics: science, services and agents on the world wide web* **2009**, *7*, 154–165.
37. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. Dbpedia: A nucleus for a web of open data. In *The semantic web*; Springer, 2007; pp. 722–735.
38. Cyganiak, R.; Wood, D.; Lanthaler, M. RDF 1.1 concepts and abstract syntax. *W3C Recommendation* **2014**, *25*, 1–8.
39. Allemang, D.; Hendler, J. *Semantic web for the working ontologist: effective modeling in RDFS and OWL*; Elsevier, 2011.
40. Brickley, D.; Guha, R.V.; McBride, B. RDF Schema 1.1. *W3C recommendation* **2014**, *25*, 2004–2014.
41. Fabbri, R. participation toolbox. https://github.com/ttm/participation, 2015.
42. Fabbri, R. social toolbox. https://github.com/ttm/social, 2015.
43. Bourahla, S.; Challal, Y. Social Networks Privacy Preserving Data Publishing. 2017 13th International Conference on Computational Intelligence and Security (CIS). IEEE, 2017, pp. 258–262.
44. Bizer, C.; Heath, T.; Berners-Lee, T. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*; IGI Global, 2011; pp. 205–227.