A Linked Open Social Dataset for Scientific Benchmarking

Renato Fabbri[a,*], Osvaldo Novais de Oliveira Junior[b]

[a] *Institute of Mathematics and Computer Sciences, University of São Paulo (ICMC/USP), Brazil*
[b] *São Carlos Institute of Physics, University of São Paulo (IFSC-USP), Brazil*

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. This work presents an open dataset with diverse provenance of human social online networking and oriented to furnish the scientific community with a friendly and common repertoire. 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 were also added as samples of less known instances. Authors recently recognized that no similar dataset was made available, and the potential benefit of the content for the research community. The linked data representation was adopted in order to comply with current best practices, homogenize access, enable discovery, and facilitate analyzes which integrate third party and provided instances. Futhermore, the method for obtaing and making the data available is potentially novel, and may of use for other research groups. This document presents an outline and overall statistics of the given dataset, which should favor subsequent work in complex and social networks, and in enhancing the available linked and social data.

Keywords: Big Data, Dataset, Benchmark Data, Facebook, Twitter, IRC, Email, Complex Networks, Social Networks, Text Mining

1. Introduction

The research on human social and complex networks often realy on data that express real phenomena [**? ? ?**]. Such data is often not made publicly available, which hinders validation of the results and does not favor makes subsequent work. Furthermore, there is a lack of open datasets for benchmarking

*Corresponding author
Email addresses: renato.fabbri@gmail.com (Renato Fabbri), chu@ifsc.usp.br (Osvaldo Novais de Oliveira Junior)

results, specially associated to the complex and social networks field, yielding diverse results from poorly related sources. Scarce and historical exceptions are the karate club or the dolphin datasets [1**?**]. For expressing the information, open machine-readable content and tools are desired for enabling data linkage and discovery [**?** **?**], and the favoured protocol is thus RDF "linked data". Social networks linked data is currently not stable or rigorous enough to be a public reference on statistical physics and social networks research, with scarce exceptions [**?** **?**] This work presents a collection of linked and open data about online human social networks, named LOSD (Linked Open Social Data). This has been achieved by linking social data into a dataset using diverse provenance, namely: Facebook, Twitter, IRC, Email groups, ParticipaBR, Cidade Democrática, and AA. Such data is proposed as a common and public repertoire for scientific research involving networks and textual content.

1.1. Related work

Most available datasets on human social networks rely on only one source [**?** **?** **?** **?**], on specialized communities such as related to research and movies [**?** **?** **?** **?** **?**], or present only statistics about social networks, and not the data that yields the networks [**?** **?** **?** **?**]. Contributions which are most similar to LOSD are mentioned in this section. In [**?**], data obtained only from Facebook is reported. With such a focus, the authors make available a number of statistics, but the data is offline at least since 2010. Linked human social network data is scarcely found in the scientific community. In [**?**], 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. [**?** **?** **?** **?**]), most often with a specific focus (users, tweets, hashtags, etc) and not presented to the scientific community by means of a research article. In summary, no human social network dataset presented to the scientific community containts the diverse provenance used in LOSD, potentially because researchers did not have the means to acquire and share the data described in Section 2.1.

2. Methods

LOSD is expressed as linked data through RDF and is ontologicaly described through the data-driven ontology synthesis method presented in Section 2.5.

2.1. Data acquisition

Research involving human subjects raise well-known ethic issues []. Furthermore, the online social networking platforms often restrict access to the data and declare illegal the use of their data through web crowlers [**?** **?**]. In this context, research using data, e.g. from Facebook, is considered problematic if not prohibited. We reached a potential solution to this issue through continuated

research with collaborators in the fields of phylosophy, social sciences, and psychology. The conceptual framework is presented in [2, 3], and in essence consists in researching first the social networks related to the researchers, and keeping the procedures, data, algorithms and results publicly available. This practices enabled collecting and sharing the data in LOSD. Notice that, although Twitter data is available through public APIs, the other sources are more problematic.

2.2. 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 to publish data on the web and 2) the use of RDF links to interlink data from different sources [**?** **?**]. The web is expected to be interconnected and to grow by the systematic application of four steps [4]:

- Use URIs to identify things [5].
- 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 [6] 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 [7, 8], and sometimes called the Giant Global Graph (GGG, akin to the WWW).

2.3. 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 [9].

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 .
```

2.4. RDFS and OWL ontologies

One of the most important notions of the semantic web is that of an ontology [**?**]. 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 [**?**]:

- 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];
- 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 [**?**]:

- 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 [**? ?**]:

- 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.

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

- 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`.

2.5. Data-driven ontology synthesis

OWL Ontologies are tools to describe taxonomies and the structure of knowledge [**?**]. Most ontologies are created by domain experts even though the data they arrange is often given by a software system and has 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 Supporting Information 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 .}
```
Such properties are used to assert existential and universal restrictions for the class.

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 query:
```
SELECT (COUNT(DISTINCT ?s) as ?cs) WHERE {
  ?s a <class_uri> .  ?s <property_uri> ?o .  ?o a ?oc .
```

```
    FILTER(str(?oc) != 'object_curi')
  }
```

When `?cs` is 0, a universal restriction is found.

4. To keep a record of further restrictions, get 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. Check if each property 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 .  }
}
```
and
```
SELECT DISTINCT ?cs WHERE { ?s <property_uri> ?o .  ?s a ?cs .  }
```

8. Render diagrams to achieve figures as exemplified in Section 4.4 and in the Supporting Information document.

3. Materials

Data was gathered from:

- public APIs (Twitter, Email);
- public logs (IRC and AA);
- Netvizz software [10] and subsequent donation by users (Facebook);
- donation by system administrators (AA, ParticipaBR, Cidade Democrática).

Integration and uniformity of access is obtained through linked data representation, as described in Section 4.5. This section introduces the underlying data in very concise terms. One should access the Supporting Information document for a more information about the dataset.

3.1. Snapshots

Of central importance to presented database 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.

3.2. Facebook data

Friendship ego networks (networks whose nodes represent friends of a user) 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.

3.3. 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.

3.4. IRC data

Public IRC (Internet Relay Chat) logs were used to obtain IRC snapshots. The database has records of relations yield by directed messages and mentions.

3.5. Email data

Email snapshots refer to individual email lists. All messages were obtained from the Gmane public email archive [11]. 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 interaction networks [12**?**]

3.6. ParticipaBR data

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

3.7. Cidade Democrática data

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

3.8. AA data

AA (Algorithmic Autoregulation [13**?**]) is a software to keep track of dedications, it is based on testifying and sharing ongoing work. The data was gathered from different versions of the system and from an IRC log.

4. Results

Current overall results concern data selection and preparation for data discovery. The main result is the data made available, which enables benchmarking of scientific results and easy experimentations, and no compliant dataset was found by the authors. Secondary results include data outline through figures and tables, software support and example SparQL queries.

4.1. Standardization

The data is embedded into URIs and triples, i.e. translated 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 the triple: `<message_id> po:text <message_text>`. Preprocessed texts are objects of triples: `<message_id> po:cleanText <message_text>`. More specialized predicates are used for delivering text 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>`. The 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; i.e. `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`.

4.2. Data outline

The database consists of 34,120,026 triples, 3,172,927 network links yield by selected interactions or relations, 382,568 participants and 253,155,020 characters in text data. Among all snapshots, 63 are ego snapshots, 54 are group snapshots; 49 have network interaction links, 89 have network friendship links; 43 have text content from messages.

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

4.3. Software tools

The authors made available software for rendering analyses and the data set itself.

4.3.1. Triplification routines

For each social platform there is a *triplification* routine, i.e. a script for expressinf the data as RDF. Original formats and further observations 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 [14]
Cidade Democrática	MySQL database	donated by admins	Participation
Email	mbox	obtained through the Gmane public archive	Gmane [11]
Facebook	GDF, GML and TAB	obtained through Netvizz [10]	Social [15]
IRC	plain text log	obtained through Supybot logging	Social
ParticipaBR	PostgreSQL database	donated by admins	Participation
Twitter	JSON	obtained through Twitter streaming API	Social

4.3.2. Topological and textual analysis

Routines are available for taking topological and textual measures from the LOSD data. Auxiliary routines, such as for performing principal component analysis (PCA) and obtaining statistics derived from the Kolmogorov-Smirnov test [**?**], are available to ease pattern recognition.

4.4. Diagrams of the data and auxiliary tables

LOSD navigation may be assisted through diagrams which describes 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.

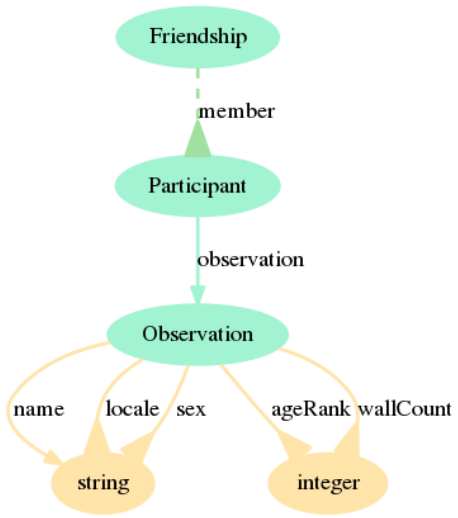

Figure 1: A diagram of the core structure that entails the friendship networks of the Facebook snapshots. A green edge denotes an OWL existential class restriction; an inverted nip denotes an OWL universal class restriction; a full (non-dashed) edge denotes an OWL functional property axiom (see Section 2.4). Further information and complete diagrams for each provenance are available in the Supporting Information document.

4.5. 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/>`.

1. Retrieve the number of participants:

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

2. 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 }
}
```

3. 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 .
}
```

4. 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
```

5. Retrieve text messages with the word "crucial" (case insensitive):

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

6. 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')
}
```

7. 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 .
}
```

8. 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 .
}
```

4.5.1. 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 the 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 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 .
}
```

Again, 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 is sufficiently similar dates, which yield networks which are not interaction or friendship networks.

5. License issues

The data presented in this article is released under public domain. Computer scripts are in git repositories and PyPI Python packages, also under public domain. Although most data is already in open licenses (Twitter, Email, Participabr, Cidade Democrática, and AA data), IRC and Facebook data was collected and donated by the individuals which yield the data, in a practice compliant with the framework described in Section 2.1.

6. Conclusions

This article outlines LOSD, a large human online social networking dataset (with almost 35 million triples) with diverse provenance. Tools and procedures used to achieve LOSD, and fundamental queries needed to retrieve the data, are also described to the convenience of the newcommer and for a consistent exposition. The resulting dataset should be expanded upon need or request from interested parties. All data should be available online in the `http://linkedopensocialdata.org` address in near future to fulfill the purpose of being a common repertoire in current research. All data is publicly available though the endpoint `xxxyyyzzz`, through which the scientific community now has consistent and curated data that enables research validation, continuity, and benchmarking. The diagrams and tables of the Supporting Information document of this article for further directions on the available structures and for an overview complement.

References

[1] M. Newman, Networks: an introduction, Oxford University Press, 2010.

[2] R. FABBRI, What are you and I? [Anthropological physics fundamentals].

[3] D. Antunes, R. Fabbri, M. M. Pisani, Anthropological physics and social psychology in the critical research of networks, International Conference on Complex Systems (2015).

[4] T. Berners-Lee, Design issues: Linked data (2006).

[5] L. Masinter, T. Berners-Lee, R. T. Fielding, Uniform resource identifier (uri): Generic syntax.

[6] J. Umbrich, S. Decker, M. Hausenblas, A. Polleres, A. Hogan, Towards dataset dynamics: Change frequency of linked open data sources.

[7] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, Dbpedia-a crystallization point for the web of data, Web Semantics: science, services and agents on the world wide web 7 (3) (2009) 154–165.

[8] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: The semantic web, Springer, 2007, pp. 722–735.

[9] R. Cyganiak, D. Wood, M. Lanthaler, Rdf 1.1 concepts and abstract syntax, W3C Recommendation 25 (2014) 1–8.

[10] B. Rieder, Studying facebook via data extraction: the netvizz application, in: Proceedings of the 5th Annual ACM Web Science Conference, ACM, 2013, pp. 346–355.

[11] R. Fabbri, gmane toolbox, https://github.com/ttm/gmane (2015).

[12] C. Bird, A. Gourley, P. Devanbu, M. Gertz, A. Swaminathan, Mining email social networks, in: Proceedings of the 2006 international workshop on Mining software repositories, ACM, 2006, pp. 137–143.

[13] R. Fabbri, R. Fabbri, V. Vieira, D. Penalva, D. Shiga, M. Mendonça, A. Negrão, L. Zambianchi, G. S. Thumé, The algorithmic autoregulation software development methodology/a metodologia de desenvolvimento de software autorregulação algorítmica, Revista Electronica de Sistemas de Informaçao 13 (2) (2014) 1.

[14] R. Fabbri, participation toolbox, https://github.com/ttm/participation (2015).

[15] R. Fabbri, social toolbox, https://github.com/ttm/social (2015).