A Linked Open Social Database for Scientific Benchmarking

Renato Fabbri[a,1,*], Osvaldo Novais de Oliveira Junior[a,1]

[a] *São Carlos Institute of Physics, São Paulo University, Brazil*

Abstract

The fields of social network analysis and complex networks are widely researched. Recently, a myriad of results have been reported which are based in diverse datasets most often not accessible to researchers other than the publishing authors. This work exposes an open database with diverse provenance and oriented to furnish the scientific community with a friendly and common repertoire. Current data was obtained from Facebook, Twitter, IRC, Email and the detached instances of ParticipaBR, AA and Cidade Democrática. These were represented as linked data to homonenize access, conform to current best practices and ease analyzes which integrate third party and provided instances. This document presents an outline and overall statistics of the given database which should favor subsequent work.

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

1. Introduction

In recent years, the web of linked data [1] has attracted wide attention in both research and application realms. However, there is a lack of datasets for benchmarking results, specially in the complex networks field, yielding diverse results from poorly related data.

*Corresponding author

Email addresses: fabbri@usp.br (Renato Fabbri), chu@ifsc.usp.br (Osvaldo Novais de Oliveira Junior)

[1] *URL:* http://www.ifsc.usp.br/

The enormity of the digital data propels a rapid development of analysis methods from different perspectives. However, the used datasets in the literature differ within the scope of each research with scarce and historical exceptions such as the karate club dataset [2]. On the other hand, the available linked data is not stable or rigorous enough to be a public reference on statistical physics and social networks research.

This work presents a linked open social database (LOSD) with diverse provenance, including Facebook, Twitter, IRC, Email and detached instances. Such data is proposed as a common repertoire for scientific research involving networks and textual content.

2. Materials

Data was gathered from:

- public APIs (Twitter, Email);
- public logs (IRC and AA);
- Netvizz software [3] 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 exposed in Section 4.5. This section introduce the underlying data in very concise terms. One should access the Supporting Information document for a thorough presentation of the database.

2.1. 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.

2.2. Facebook data

Friendship ego networks (networks whose constituents are friends of an 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.

2.3. Twitter data

Tweets were gathered through the Twitter streaming public API. Each snapshot is unified by a distinct hashtag. Edges are canonically yield by retweets but replies and user mentions are also kept in the database.

2.4. IRC data

Public IRC logs were used to render IRC snapshots. The database has records of users to which the message is directed to or mentions.

2.5. Email data

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

2.6. ParticipaBR data

The ParticipaBR is a Brazilian federal platform for social participation. Texts are derived from blog posts and networks are derived from friendship and interaction criteria.

2.7. Cidade Democrática data

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

2.8. AA data

The Algorithmic Autoregulation [5] is a software development methodology based on testifying and sharing ongoing work. The data was gathered from different versions of the system and from an IRC log.

3. Methods

Data is represented as linked open data through RDF and ontologicaly described through a data-driven ontology synthesis method.

3.1. 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 web of data is constructed with documents on the web such as the web of HTML documents. In practice, the idea 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 [1]:

- Use URIs to identify things [6].
- Use HTTP URIs.
- Provide useful information when an 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 [7] builds an ever 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 [8, 9].

3.2. 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, all which, in essence, represent a labeled and directed multi-graph. RDF may be stored in a type of database called a triplestore [10].

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

```
http://example.org/Things#Paper http://example.org/hasColor
http://example.org/Colors#White .
```

3.3. Data-driven ontology synthesis

OWL Ontologies are critical 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 data with SPARQL queries and post-processing. The results are OWL code and diagrams which are available in the Supporting Information document. 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`). If the numbers match, there is an existential restriction for the class. The queries are:

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

4. Find the number of instances which are subjects of triples where the predicate is the property but are not instances of the class. If there is zero of such instances, there is an universal restriction:

```
SELECT (COUNT(DISTINCT ?s)=0 as ?cs) WHERE {
  ?s <property_uri> ?o .  ?s a ?ca .  FILTER(str(?ca) != 'class_uri')
}
```

5. To keep a record of the restrictions (and occurring triples), 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 .  }
}
```

6. Obtain all distinct properties:

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

7. 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
```

8. For each property, find the incident range and domain 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 .  }
```

9. Render diagrams as exposed in the next section and in the Supporting Information file.

4. Results

Current overall results concern data selection and preparation for knowledge discovery. The main result is the data made available, which enables benchmarking of scientific results and easy experimentations. Secondary results include data outline through figures and tables, software support and example SparQL queries.

4.1. Standardization

The data is embedded into standard 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.

Instances are built on top of the class they derive from plus a hashtag character, a provenance string (e.g. `facebook-legacy` or `participabr-legacy`) of the snapshot they refer to, 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 edges yield by interactions or relations, 382,568 participants and 253,155,020 characters. Among all snapshots, 63 are ego snapshots, 54 are group snapshots; 49 have interaction edges, 89 have friendship edges; 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 database is released with software for rendering itself, analyses and multimedia artifacts.

4.3.1. Triplification routines

For each social platform there is a *triplification* routine, i.e. a script for translating data to 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 [11]
Cidade Democrática	MySQL database	donated by admins	Participation
Email	mbox	obtained through Gmane public database	Gmane [4]
Facebook	GDF, GML and TAB	obtained through Netvizz [3]	Social [12]
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 database. Auxiliary routines, such as performing principal component analysis and taking Kolmogorov-Smirnov measures, are available to ease pattern recognition. Single, timeline and multi-scale analyzes are automated.

4.3.3. Multimedia rendering

It is a core purpose of framework to provide routines for rendering audiovisualizations of the data. Social structures are rendered into music, images and video animations through the Percolation toolbox [13] in association with the Music and Visuals toolboxes [14, 15].

4.3.4. Migration from deprecated toolboxes

Routines mentioned in Sections 4.3.2 and 4.3.3 are being migrated from deprecated toolboxes [16, 17] into newly designed toolboxes [13, 15].

4.4. Diagrams of the data and auxiliary tables

The database exploration can be assisted through diagrams which expose the structure from each provenance. Such diagrams are in the Supporting Information document with some tables to ease understanding of the provided data. A simplified example is given in Figure 1 where the friendship structure of the Facebook snapshots are exposed.

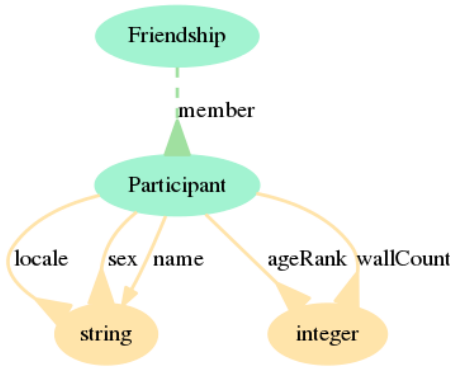

Figure 1: A diagram of the structure involved in the friendship networks of the Facebook snapshot. A green edge denotes an OWL existential class restriction; an inverted nip denotes an OWL universal class restriction; a full (non-dashed) edge denotes an OWL functional property axiom. Further information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

4.5. SPARQL queries

There are numerous useful and general purpose SPARQL queries to be performed against the database. Here we write some of the most basic of such queries selected by their potential to be varied. All queries assume the use 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 an specific user:

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

5. Search string "pineapple" in LOSD messages:

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

5. License issues

The database 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. This rises the the understanding of the right to study such data as the right to access the self, in parity with anthropological endaviors [18, 19].

6. Conclusions

The database presented in this article constitutes a large database with diverse provenance. Even so, database should be expanded in upon need or requests from feedback. 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. One should reach 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] T. Berners-Lee, Design issues: Linked data (2006).

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

[3] 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.

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

[5] 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.

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

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

[8] 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.

[9] 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.

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

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

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

[13] R. Fabbri, Percolation toolbox, https://github.com/ttm/percolation (2015).

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

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

[16] R. Fabbri, Gmane legacy repository, https://github.com/ttm/gmaneLegacy (2015).

[17] R. Fabbri, Percolation legacy repository, https://github.com/ttm/percolationLegacy (2015).

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

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