Linked Open Social Data 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 field of social network analysis and the topic of complex networks are widely researched. Recently, a myriad of results have been reported which are based in diverse datasets most often not accessible to other researchers. This work exposes an open dataset with diverse provenance and oriented to provide the scientific community 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 dataset which should favor subsequent work.

Keywords: Benchmark Data, Facebook, Twitter, IRC, Email, Complex Networks

1. Introduction

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

The enormity of the digital data propels a rapid development of analysis methods from different perspectives. The used datasets differ within the scope

*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/

of each research with scarce and historical exception 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 data (LOSD) dataset with data from diverse provenance, including Facebook, Twitter, IRC, Email and detached instances. Such data is proposed as a common repertoire for scientific reseach involving networks and textual content.

2. Materials

Data was gathered either from:

- public APIs (Twitter, Email); or
- public logs (IRC and AA); or
- Netvizz software [3] (Facebook); or
- donated data from users (Facebook); or
- donated data from system administrators (AA, ParticipaBR, Cidade Democrática).

Integration and uniformity of access is obtained through linked data representation, as exposed in Section 4.3.

Of central importance to presented LOSD is the concept of a snapshot. A snapshot is 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 event is a snapshot; the friendship, interaction and posts structures of a facebook group, prospected at the same time, is a snapshot.

2.1. Facebook data

Friendship ego networks (networks whose reference is 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.

2.2. Twitter data

Tweets were gathered through the 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 LOSD.

2.3. IRC data

Public IRC logs were used to render LOSD IRC snapshots. LOSD has record of users to which the message is directed to or mentions.

2.4. Email data

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

2.5. ParticipaBR data

The ParticipaBR is a platform for social participation once regarded as the Brazilian portal of social participation. Texts are derived from blog posts and networks are derived from friendship and interaction criteria.

2.6. AA data

The Algorithmic Autoregulation [5] is a methodology for testifying and sharing ongoing work. The data was gathered from different versions of the system and from IRC logs and is presented as part of the LOSD as one of the detached platforms.

2.7. Cidade Democrática data

Cidade Democrática is a civil society social participation portal.

3. Methods

Data in the presented LOSD is represented as linked open data through RDF and ontologicaly described through a data-driven ontology synthesis method. These steps are described in the following sections.

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 hypertext. 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/paper http://example.org/hasColor
http://example.org/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.

We developed an ontology synthesis method that probes the ontological structure in data with SPARQL queries and post-processing which can be divided in the following steps:

1. Obtaining all distinct classes with the query:
   ```
   SELECT DISTINCT ?class WHERE { ?s a ?class }
   ```

2. Obtaining all distinct properties with the query:
   ```
   SELECT DISTINCT ?p WHERE { ?s ?p ?o }
   ```

3. For each class, get distinct subject classes and predicates where the the object is an instance of the class:
   ```
   SELECT DISTINCT ?p ?cs WHERE { ?i a <class_uri> .  ?s ?p ?i .  ?s a
   ?cs .  }
   ```

4. For each class, get distinct predicates and object classes or datatypes where the subject is an instance of such class:
   ```
   SELECT DISTINCT ?p ?co (datatype(?o) as ?do) WHERE { ?i a <class_uri>
   .  ?i ?p ?o .  OPTIONAL { ?o a ?co .  } }
   ```

5. For each property, check if it is functional, i.e. if it occurs only once with each subject:
   ```
   SELECT DISTINCT (COUNT(?o) as ?co) WHERE { ?s <property_uri> ?o } GROUP
   BY ?s
   ```

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

7. For each instance of each class, get all distinct predicates. For each predicate, check if all instances of the class hold such relationship (existential restriction):
   `SELECT DISTINCT ?p WHERE { ?s a <class_uri>. ?s ?p ?o . }`
   `SELECT DISTINCT ?s WHERE { ?s a <class_uri> }`
   `SELECT DISTINCT ?s ?co (datatype(?o) as ?do) WHERE {?s a <class_uri>. ?s <property_uri> ?o . OPTIONAL {?o a ?co . }}`

8. and if all instances that hold such relationship are instances of the class (universal restriction):
   `SELECT DISTINCT ?s WHERE { ?s <property_uri> ?o . }`

9. Draw each class, each property and the overall figure.

10. Make `rdfs:subClassOf` and `rdfs:subPropertyOf` statements to better organize knowledge and link to third party ontologies and data.

4. Results

4.1. Data outline

4.2. Software tools

4.3. SPARQL queries

5. Conclusions

The Linked Open Social Data (LOSD) presented in this article 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.

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] L. M. Ingebrigtsen, Gmane (2008).

[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.