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 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 dataset which should favor subsequent work.

Keywords: Big Data, Data Mining, 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 a lack of datasets for research benchmarking, 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 data (LOSD) dataset with 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 by users (Facebook); or
- donated data by system administrators (AA, ParticipaBR, Cidade Democrática).

Integration and uniformity of access is obtained through linked data representation, as exposed in Section 4.4.

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 .  } }
   ```

   and

   ```
   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

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 software support and SparQL queries made beforehand.

4.1. Standardization

The LOSD is a collection of data translated into RDF and embedded into standard URIs and triples. Such standards constitute the LOSD itself, but some of the conventions are outlined in this section.

URIs have the root `http://purl.org/socialparticipation/participationontology/` which are identified with the radical `po:`. Classes and properties are build by adding a suffix to the root, as in `po:Participant` or `po:text`. Classes have "UpperCamelCase" suffixes while properties have "lowerCamelCase" suffixes. All participants and transactions (messages, friendships, interactions) are linked to snapshots through the triple `<resource> po:snapshot <snapshot_uri>`. Message texts, including comments, are objectives in the triple: `<message_id> po:text <message_text>`. Individuals (also called instances) are built on top of the class they derive from plus a hashtag character, the snapshot id of the snapshot they refer to, and an identifier; i.e. `po:Participant#<snapshot_id>-<local_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-c++...`.

4.2. Data outline

Database consists of 37467423 triples, 3088219 edges yield by interactions or relations, 382735 participants and 229008584 characters. Among all snapshots, 63 are ego snapshots, 53 are group snapshots; 45 have interaction edges, 89 have friendship edges; 42 have text content from messages.

Table 1: Number of triples (ntriples), number of relations/interactions/edges (nedges), number of participants (nparticipants) and number of characters (nchars) in each LOSD snapshot.

snapshot id	**ntriples**	**nedges**	**nparticipants**	**nchars**
aa-irc-legacy-foradoeixo.log	22	0	1	0
aa-irc-legacy-hackerspace-cps.log	36	0	2	0
aa-irc-legacy-hackerspaces-br.log	64	0	5	0
aa-irc-legacy-labmacambira_lalenia.txt	53067	0	117	0
aa-mongo-legacy	22773	0	37	0
aa-mysql-legacy	790796	0	157	0
cidadedemocratica-legacy	1548412	0	23079	0
facebook-legacy-AdornoNaoEhEnfeite29032013	8841	1292	293	26113
facebook-legacy-AntonioAnzoategui18022013	1852	328	52	0
facebook-legacy-AtivistasDaInclusaoDigital09032013	25642	5592	306	0

Continued on next page

Table 1 – *Continued from previous page*

snapshot id	ntriples	nedges	nparticipants	nchars
facebook-legacy-Auricultura10042013	19905	3898	412	14015
facebook-legacy-BrunoMialich31012013	40970	9320	502	0
facebook-legacy-CalebLuporini13042013	106138	24653	1050	0
facebook-legacy-CalebLuporini19022013	104922	24391	1026	0
facebook-legacy-CienciasComFronteiras29032013	110734	23302	2921	0
facebook-legacy-ComputerArt10032013	260050	62819	1342	0
facebook-legacy-Coolmeia06032013	76063	16534	1202	0
facebook-legacy-DanielPenalva18022013	3695	682	113	0
facebook-legacy-DemocraciaDiretaJa14032013	259061	59323	3053	54443
facebook-legacy-DemocraciaDiretaJa14072013	4162	310	214	58035
facebook-legacy-DemocraciaPura06042013	32627	6730	627	65062
facebook-legacy-Economia14042013	239190	54001	3587	52664
facebook-legacy-EconomiaCriativaDigital03032013	185905	43128	1684	0
facebook-legacy-EducacoesEAprendizagensXXI02032013	106918	24802	1285	0
facebook-legacy-GabrielaThume19022013	18757	4108	307	0
facebook-legacy-GrahamForrest28012013	1546	185	90	0
facebook-legacy-LailaManuelle17012013	201247	48572	969	0
facebook-legacy-LarissaAnzoategui20022013	25000	5191	580	0
facebook-legacy-Latesfip08032014	11554	2009	306	0
facebook-legacy-LivingBridgesPlanet29032013	150032	32494	3077	52808
facebook-legacy-LuisCirne07032013	16795	3390	437	0
facebook-legacy-MariliaMelloPisani10042013	84946	19040	1230	0
facebook-legacy-Mirtes16052013	39591	9075	445	0
facebook-legacy-MobilizacoesCulturaisInteriorSP13032013	26518	6096	298	0
facebook-legacy-PartidoPirata23032013	45883	8537	1943	36313
facebook-legacy-PedroPauloRocha10032013	216064	50591	1932	0
facebook-legacy-PeterForrest28012013	8332	1829	120	0
facebook-legacy-PoliticasCulturasBrasileiras08032013	178689	41690	1278	69756
facebook-legacy-PracaPopular16032013	4577	932	65	4249
facebook-legacy-RafaelReinehr09042013	174599	39586	2297	0
facebook-legacy-RamiroGiroldo20022013	10104	2020	264	0
facebook-legacy-RedeTranzmidias02032013	25497	4940	391	54907
facebook-legacy-RenatoFabbri02032013	93310	21579	974	0
facebook-legacy-RenatoFabbri03032013	93866	21711	978	0
facebook-legacy-RenatoFabbri11072013	114728	26440	1256	0
facebook-legacy-RenatoFabbri18042013	104332	24072	1124	0
facebook-legacy-RenatoFabbri20012013	86592	20085	868	0
facebook-legacy-RenatoFabbri29112012	82269	19083	823	0
facebook-legacy-RicardoFabbri18022013	11548	2327	344	0
facebook-legacy-RitaWu08042013	84071	18935	1165	0
facebook-legacy-RonaldCosta12062013	31514	6557	730	0

Continued on next page

Table 1 – *Continued from previous page*

snapshot id	ntriples	nedges	nparticipants	nchars
facebook-legacy-SiliconValleyGlobalNetwork27042013	77552	15740	2130	50251
facebook-legacy-SolidarityEconomy12042013	14614	2404	525	67774
facebook-legacy-StudyGroupSNA05042013	5856	480	448	25474
facebook-legacy-THackDay26032013	1844	420	41	0
facebook-legacy-Tecnoxamanismo08032014	11106	2069	318	0
facebook-legacy-Tecnoxamanismo15032014	14589	2702	450	0
facebook-legacy-ThaisTeixeira19022013	26600	6088	296	0
facebook-legacy-VilsonVieira18022013	19864	4334	336	0
facebook-legacy-ViniciusSampaio18022013	90691	21360	725	0
facebook-legacy-avlab_BarthorLaZule22022014	16005	3513	279	0
facebook-legacy-avlab_CalebLuporini25022014	125577	29268	1215	0
facebook-legacy-avlab_CamilaBatista23022014	21138	4476	462	0
facebook-legacy-avlab_CarlosDiego25022014	171744	39401	2020	0
facebook-legacy-avlab_CristinaMekitarian23022014	24647	5572	337	0
facebook-legacy-avlab_DanielGonzales23022014	196406	45318	2162	0
facebook-legacy-avlab_FelipeBrait23022014	1228605	299082	4611	0
facebook-legacy-avlab_FelipeVillela22022014	2475	477	81	0
facebook-legacy-avlab_JoaoMeirelles25022014	52371	11649	825	0
facebook-legacy-avlab_JoaoMekitarian23022014	88765	20821	783	0
facebook-legacy-avlab_JulianaSouza23022014	129757	29942	1427	0
facebook-legacy-avlab_KarinaGomes22022014	9073	1906	207	0
facebook-legacy-avlab_LucasOliveira26022014	62871	14764	545	0
facebook-legacy-avlab_MarcelaLucatelli25022014	138733	31647	1735	0
facebook-legacy-avlab_MariliaPisani25022014	114765	25830	1635	0
facebook-legacy-avlab_NatachaRena22022014	642769	154758	3391	0
facebook-legacy-avlab_OrlandoCoelho22022014	5149	848	251	0
facebook-legacy-avlab_PalomaKliss25022014	493774	119520	2242	0
facebook-legacy-avlab_PedroRocha25022014	346883	81910	2749	0
facebook-legacy-avlab_RenatoFabbri22022014	124703	28780	1369	0
facebook-legacy-avlab_SarahLuporini25022014	505853	121502	2835	0
facebook-legacy-avlab_SatoBrasil25022014	1519394	371249	4914	0
facebook-legacy-ego_MarceloSaldanha19112014	130556	29440	1828	0
facebook-legacy-ego_MariliaPisani06052014	122231	27581	1701	0
facebook-legacy-ego_MassimoCanevacci19062013	273328	59995	4764	0
facebook-legacy-ego_RenatoFabbri06022014	123993	28606	1367	0
facebook-legacy-ego_RenatoFabbri19112014	153410	35514	1622	0
facebook-legacy-ego_VJPixel23052014	231608	54752	1800	0
facebook-legacy-posavlab_AnaCelia18032014	53939	12167	753	0
facebook-legacy-posavlab_ElenaGarnelo04032014	93472	21723	940	0
facebook-legacy-posavlab_FabiBorges08032014	159584	36592	1888	0
facebook-legacy-posavlab_GeorgeSanders08032014	108071	24706	1321	0

Continued on next page

Table 1 – *Continued from previous page*

snapshot id	ntriples	nedges	nparticipants	nchars
facebook-legacy-posavlab_GrazielleMachado18032014	24264	5254	464	0
facebook-legacy-posavlab_RenatoFabbri19032014	129360	29890	1400	0
facebook-legacy-posavlab_RicardoPoppi18032014	76104	17234	1024	0
gmane-legacy-.comp.gcc.libstdc++.devel1-20000	364045	14786	1036	30126252
gmane-legacy-.linux.audio.devel1-20000	418957	17076	1232	26969596
gmane-legacy-.linux.audio.users1-20000	390944	16362	1147	25065928
gmane-legacy-.politics.organizations.metareciclagem1-20000	378641	15230	477	54260954
irc-legacy-foradoeixo	859401	4308	3318	3023060
irc-legacy-hackerspace-cps	286886	1860	607	454655
irc-legacy-hackerspaces-br	1347458	210	347	8029920
irc-legacy-labmacambira	1964597	58525	1561	6535187
participabr-legacy	198157	0	3825	0
twitter-legacy-ChennaiFloods	7255622	86935	46493	23237802
twitter-legacy-ForaCunha	95613	1406	2747	372131
twitter-legacy-ForaDilma	28554	534	659	113810
twitter-legacy-MAMA2015	21736361	411971	33080	75358785
twitter-legacy-QuartaSemRacismoClubeSDV	395019	4023	5785	1635867
twitter-legacy-SnapDetremura	28896	405	621	124448
twitter-legacy-arenaNETmundial	414868	13291	5898	2825121
twitter-legacy-art	2980883	26501	30486	9539413
twitter-legacy-dilma	83663	1563	2274	332005
twitter-legacy-fuck	3872	30	93	14727
twitter-legacy-game	245251	1370	4682	1229910
twitter-legacy-god	1613307	17861	22117	5560140
twitter-legacy-music	7811979	31642	51617	19540393
twitter-legacy-obama	1220179	17080	20330	4481840
twitter-legacy-porn	1303814	4935	5941	4970218
twitter-legacy-python	5852	3	17	1758
twitter-legacy-science	394491	3312	7156	1910216

4.3. Software tools

The LOSD is released with a 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: Number of snapshots from each provenance. Every snapshot is a `po:Snapshot`; there are three types of the `po:AASnapshot` class.

snapshot provenance	**number of snapshots**
http://purl.org/socialparticipation/po/AAIRCSnapshot	5
http://purl.org/socialparticipation/po/AAMongoSnapshot	1
http://purl.org/socialparticipation/po/AAMysqlSnapshot	1
http://purl.org/socialparticipation/po/AASnapshot	7
http://purl.org/socialparticipation/po/CidadeDemocraticaSnapshot	1
http://purl.org/socialparticipation/po/FacebookSnapshot	88
http://purl.org/socialparticipation/po/IRCSnapshot	4
http://purl.org/socialparticipation/po/ParticipabrSnapshot	1
http://purl.org/socialparticipation/po/TwitterSnapshot	16
http://purl.org/socialparticipation/po/Snapshot	116

Table 3.

Table 3: Social platforms, original formats and further observations for the LOSD dataset.

social platform	**original format**	**further observations**	**toolbox**
AA	MySQL and MongoDB databases; IRC text logs	donated by AA users	Participation
Cidade Democrática	MySQL database	donated by admins	Participation
Email	mbox	obtained through Gmane public database	Gmane
Facebook	GDF, GML and TAB	obtained through Netvizz [3]	Social
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 dataset. Principal Component Analysis (PCA) is performed with available measures to ease pattern recognition both in a timeline and in a multiscale fashion.

4.3.3. Multimedia rendering

It is a core purpose of LOSD framework to provide routines for rendering audiovisualizations of the LOSD data. Social structures are rendered into music, images and video animations through the Percolation toolbox [11] in association with the Music and Visuals toolboxes [12, 13].

4.3.4. Migration from deprecated toolboxes

Routines mentioned in Sections 4.3.2 and 4.3.3 are being migrated from deprecated toolboxes [14, 15] into newly designed toolboxes [11, 13].

4.4. SPARQL queries

There are numerous useful and general purpose SPARQL queries to be performed agaist the LOSD 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 preable `PREFIX po: <http://purl.org/socialparticipation/po/>`.

1. Retrieve the number of participants
2. Retrieve the number of relations, be them interactions of friendships
3. Retrieve the number of relations, be them interactions of friendships
4. Retrieve all text produced by an specific user
5. List 100 users with the most friendships
6. List 100 users with most interactions
7. Search string in messages

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. One should access `http://wiki.nosdigitais.teia.org.br/LOSD` for reaching the address where LOSD is currently reachable.

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.

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

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

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

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

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