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

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
twitter-legacy-arttw.pickle	2866292	26501	30486	9539413
StudyGroupSNA05042013.gdf_fb	5661	480	448	25474
SiliconValleyGlobalNetwork27042013.gdf_fb	77194	15740	2130	50251
THackDay26032013.gdf_fb	1844	420	41	[]
RitaWu08042013.gml_fb	83752	18935	1165	[]
legacy-gmane.linux.audio.devel-1-20000	418949	17076	1232	26969596
BrunoMialich31012013.gml_fb	40754	9320	502	[]

Continued on next page

Table 1 – *Continued from previous page*

snapshot id	ntriples	nedges	nparticipants	nchars
GabrielaThume19022013.gml_fb	18565	4108	307	[]
CalebLuporini13042013.gml_fb	105861	24653	1050	[]
RedeTranzmidias02032013.gdf_fb	24950	4940	391	54907
twitter-legacy-godtw.pickle	1534285	17861	22117	5560140
avlab_KarinaGomes22022014.gdf_fb	9073	1906	207	[]
posavlab_RenatoFabbri19032014.gdf_fb	129338	29890	1400	[]
GrahamForrest28012013.gml_fb	1362	185	90	[]
RamiroGiroldo20022013.gml_fb	9911	2020	264	[]
Coolmeia06032013.gdf_fb	75875	16534	1202	[]
ComputerArt10032013.gdf_fb	259924	62819	1342	[]
DemocraciaDiretaJa14032013.gdf_fb	258541	59323	3053	54443
AdornoNaoEhEnfeite29032013.gdf_fb	8416	1292	293	26113
avlab_MariliaPisani25022014.gdf_fb	114752	25830	1635	[]
avlab_CalebLuporini25022014.gdf_fb	125568	29268	1215	[]
avlab_LucasOliveira26022014.gdf_fb	62870	14764	545	[]
RonaldCosta12062013.gml_fb	31267	6557	730	[]
legacy-gmane.politics.organizations.metareciclagem-1-20000	378633	15230	477	54260954
ego_MarceloSaldanha19112014.gdf_fb	130545	29440	1828	[]
LailaManuelle17012013.gml_fb	201004	48572	969	[]
CalebLuporini19022013.gml_fb	104649	24391	1026	[]
twitter-legacy-MAMA2015tw.pickle	20870910	411971	33080	75358785
CienciasComFronteiras29032013.gdf_fb	110728	23302	2921	[]
posavlab_RicardoPoppi18032014.gdf_fb	76098	17234	1024	[]
avlab_JulianaSouza23022014.gdf_fb	129736	29942	1427	[]
Auricultura10042013.gdf_fb	273206	60088	412	14015
DanielPenalva18022013.gml_fb	3507	682	113	[]
RicardoFabbri18022013.gml_fb	11344	2327	344	[]
AntonioAnzoategui18022013.gml_fb	1676	328	52	[]
ego_RenatoFabbri06022014.gdf_fb	123973	28606	1367	[]
RenatoFabbri03032013.gml_fb	93599	21711	978	[]
Mirtes16052013.gml_fb	39384	9075	445	[]
avlab_JoaoMekitarian23022014.gdf_fb	88764	20821	783	[]
avlab_PalomaKliss25022014.gdf_fb	493740	119520	2242	[]
RenatoFabbri20012013.gml_fb	86338	20085	868	[]
avlab_DanielGonzales23022014.gdf_fb	196330	45318	2162	[]
PeterForrest28012013.gml_fb	8145	1829	120	[]
PedroPauloRocha10032013.gml_fb	215647	50591	1932	[]
twitter-legacy-obamatw.pickle	1169258	17080	20330	4481840
avlab_BarthorLaZule22022014.gdf_fb	16005	3513	279	[]
twitter-legacy-sciencetw.pickle	374480	3312	7156	1910216
twitter-legacy-porntw.pickle	1240228	4935	5941	4970218

Continued on next page

Table 1 – *Continued from previous page*

snapshot id	ntriples	nedges	nparticipants	nchars
legacy-gmane.linux.audio.users-1-20000	390936	16362	1147	25065928
RenatoFabbri18042013.gml_fb	104048	24072	1124	[]
posavlab_GeorgeSanders08032014.gdf_fb	108029	24706	1321	[]
avlab_MarcelaLucatelli25022014.gdf_fb	138694	31647	1735	[]
avlab_SarahLuporini25022014.gdf_fb	505835	121502	2835	[]
Tecnoxamanismo15032014.gdf_fb	14406	2702	450	[]
avlab_FelipeVillela22022014.gdf_fb	2475	477	81	[]
EducacoesEAprendizagensXXI02032013.gdf_fb	106894	24802	1285	[]
Latesfip08032014.gdf_fb	11212	2009	306	[]
MariliaMelloPisani10042013.gml_fb	84691	19040	1230	[]
avlab_OrlandoCoelho22022014.gdf_fb	5143	848	251	[]
irc-legacy-labmacambira_lalenia.txt	1960692	58521	1561	6534316
posavlab_GrazielleMachado18032014.gdf_fb	97056	21016	464	[]
DemocraciaPura06042013.gdf_fb	32252	6730	627	65062
avlab_CarlosDiego25022014.gdf_fb	171724	39401	2020	[]
LuisCirne07032013.gml_fb	16588	3390	437	[]
avlab_SatoBrasil25022014.gdf_fb	1519337	371249	4914	[]
avlab_JoaoMeirelles25022014.gdf_fb	52359	11649	825	[]
avlab_RenatoFabbri22022014.gdf_fb	124681	28780	1369	[]
avlab_FelipeBrait23022014.gdf_fb	1228463	299082	4611	[]
avlab_CristinaMekitarian23022014.gdf_fb	24646	5572	337	[]
ThaisTeixeira19022013.gml_fb	26411	6088	296	[]
posavlab_ElenaGarnelo04032014.gdf_fb	93432	21723	940	[]
irc-legacy-hackerspace-cps.log	725488	3716	607	907517
RenatoFabbri29112012.gml_fb	82017	19083	823	[]
avlab_PedroRocha25022014.gdf_fb	346801	81910	2749	[]
RafaelReinehr09042013.gml_fb	174221	39586	2297	[]
twitter-legacy-SnapDetremuratw.pickle	27135	405	621	124448
RenatoFabbri02032013.gml_fb	93044	21579	974	[]
posavlab_FabiBorges08032014.gdf_fb	159515	36592	1888	[]
ego_MariliaPisani06052014.gdf_fb	122218	27581	1701	[]
ego_MassimoCanevacci19062013.gdf_fb	273237	59995	4764	[]
LarissaAnzoategui20022013.gml_fb	24779	5191	580	[]
Tecnoxamanismo08032014.gdf_fb	10979	2069	318	[]
avlab_CamilaBatista23022014.gdf_fb	21132	4476	462	[]
twitter-legacy-ForaDilmatw.pickle	27297	534	659	113810
PartidoPirata23032013.gdf_fb	45419	8537	1943	36313
PracaPopular16032013.gdf_fb	4522	932	65	4249
EconomiaCriativaDigital03032013.gdf_fb	185682	43128	1684	[]
posavlab_AnaCelia18032014.gdf_fb	53935	12167	753	[]
RenatoFabbri11072013.gml_fb	114430	26440	1256	[]

Continued on next page

Table 1 – *Continued from previous page*

snapshot id	ntriples	nedges	nparticipants	nchars
Economia14042013.gdf_fb	238649	54001	3587	52664
LivingBridgesPlanet29032013.gdf_fb	149675	32494	3077	52808
DemocraciaDiretaJa14072013.gdf_fb	257151	59781	3607	58035
avlab_NatachaRena22022014.gdf_fb	642698	154758	3391	[]
PoliticasCulturasBrasileiras08032013.gdf_fb	178132	41690	1278	69756
VilsonVieira18022013.gml_fb	19662	4334	336	[]
ego_VJPixel23052014.gdf_fb	231582	54752	1800	[]
ego_RenatoFabbri19112014.gdf_fb	153387	35514	1622	[]
irc-legacy-hackerspaces-br.log	1347450	210	347	8029920
SolidarityEconomy12042013.gdf_fb	14302	2404	525	67774
AtivistasDaInclusaoDigital09032013.gdf_fb	25542	5592	306	[]
legacy-gmane.comp.gcc.libstdc++.devel-1-20000	364037	14786	1036	30126252
MobilizacoesCulturaisInteriorSP13032013.gdf_fb	26508	6096	298	[]
irc-legacy-foradoeixo.log	2070970	8442	3318	5842836
ViniciusSampaio18022013.gml_fb	90463	21360	725	[]
twitter-legacy-QuartaSemRacismoClubeSDVtw.pickle	371328	4023	5785	1635867

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.