Unevenness in network properties on the social Semantic Web

Raf Guns

University of Antwerp, Informatie- en Bibliotheekwetenschap,
CST, Venusstraat 35, 2000 Antwerpen
raf.guns@ua.ac.be

Abstract. This paper studies unevenness in network properties on the social Semantic Web. First, we propose a two-step methodology for processing and analyzing social network data from the Semantic Web, based on the SPARQL query language. After a brief introduction to the notion of unevenness, the methodology is applied to examine unevenness in network properties of real-world data. Comparing Lorenz curves for different centrality measures, it is shown how examinations of unevenness can provide crucial hints regarding the topology of (social) Semantic Web data.

Key words: social network analysis, Semantic Web, SPARQL, unevenness

1 Introduction

The *social Semantic Web* is a broad, non-technical term, referring to data on the Semantic Web (encoded in RDF) that contain social information. The most prevalent ontology on the social Semantic Web is the FOAF (Friend Of A Friend) vocabulary [8]. Yet, FOAF is not alone; in this paper, for instance, we will use a socio-cultural ontology (section 4).

The Semantic Web [5] in general is conceived as a large-scale distributed information system. While some constituents are still in development and its current uptake is relatively modest, the Semantic Web graph already shows the traits of a complex system. As such, it is characterized by [3, 15]:

Skewed degree distribution: The probability $P(k)$ that a node has degree k (is connected to k other nodes) is not randomly distributed. Instead, it follows a power law $P(k) \approx Ak^{-\gamma}$. Moreover, complex systems typically exhibit power law distributions in more than one way. With regard to the Semantic Web, previous research has shown that a diversity of relations — such as the relation between websites and their number of Semantic Web documents or the relation between an ontology and its number of uses — follows a power law [13].

Small world properties: Made famous by Stanley Milgram's [20] letter experiment, the small world notion refers to the fact that the average shortest path

length in a graph is very short (comparable to that of a random graph). More recently, several models have been proposed to account for the small-world effect [21, 27].

High clustering: The neighbours of a given node are likely also neighbours of each other.

Similar traits have been discovered for a variety of social and biological networks [10]. However, these properties also raise several questions. In this paper, we will address two of them. Both questions will be discussed and demonstrated on a real-world socio-cultural data set.

First, how can data on the social Semantic Web be used for Social Network Analysis (SNA)? Significant research in this area has already been performed by, among others, Li Ding and colleagues [12] and Peter Mika [19]. Much work has concentrated on acquiring and aggregating data (often FOAF data), – especially merging information about unique persons turns out to be far from trivial. In the present paper, we concentrate on the development of a methodology for using one single RDF graph as the 'master', which can be used as the basis for several kinds of SNA. Ideally, we want to keep as much information as possible and extract a multitude of potentially interesting relations. This particular aspect has received less attention so far.

Second, it is very rarely examined *how* skewed a distribution is. How can this notion be measured? Quantification of unevenness is crucial for a thorough understanding of a power law distribution; moreover, it can be used for comparison purposes between distributions and between networks.

2 Two-step methodology

Semantic Web data can be stored in many different ways: as a (set of) document(s) in one of the many RDF syntaxes [4]; in a 'classic' relational database; or in a *triplestore*, a dedicated RDF database. For the remainder of this paper, we assume the use of a triplestore (see [17] for an overview of triplestores), using Jena[1] as an example. Triplestores can be queried with a query language like SPARQL [23].

Partly due to its distributed nature, Semantic Web data may appear quite dazzling: many different kinds of data, drawn from several ontologies, between which a multitude of relations exist. How can one make heads or tails out of them?

Assuming the existence of a set of fairly clearly defined questions to be answered, we propose a two-step methodology, which critically depends on SPARQL (or a query language with similar capabilities). In short, the two steps are:

[1] Internally, Jena uses a relational database, but the interface is similar to other triplestores, see `http://jena.sourceforge.net/DB/creating-db-models.html`.

1. Construct an *extraction query* in SPARQL and apply it to the RDF graph. This yields a secondary graph, specifically oriented towards the question(s).
2. Convert the secondary graph to a format intended for SNA.

We will now discuss both steps in greater detail.

2.1 Constructing an extraction query

SPARQL queries are usually `SELECT` queries, which return a table of results. For the extraction query, we employ `CONSTRUCT` queries, which return a new RDF graph. A similar architecture can also be found in the MESUR project [7, 24].

First, we compare the original graph in the triplestore and the questions to be answered. Some questions simply involve the *extraction* of parts of the RDF graph (ignoring the rest). A typical example would be the extraction of all *foaf:knows* relations from a FOAF triplestore. This can actually be done without SPARQL, but for the sake of illustration we give a possible extraction query:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

CONSTRUCT { ?p1 foaf:knows ?p2 }
WHERE {
  ?p1  a           foaf:Person ;
       foaf:knows  ?p2 .
  ?p2  a           foaf:Person .
}
```

Other questions are trickier, in that they require knowledge on how relations in the model interact, — these involve *extraction and combination* of parts of the model. Let's use the IngentaConnect MetaStore project [22], a large-scale database of academic articles, as an example. Fig. 1 shows how article citations are expressed in MetaStore. The citation relation *between authors* can then be queried as follows.

```
BASE <http://metastore.ingentaconnect.com>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX prism: <http://prismstandard.org/namespaces/1.2/basic>
PREFIX ex: <http://example.com/ns/>

CONSTRUCT { ?author1 ex:cites ?author2 }
WHERE {
  ?art1  a                </ns/structure/Article> ;
         foaf:maker       ?author1 ;
         prism:references ?art2 .
  ?art2  a           </ns/structure/Article> ;
         foaf:maker  ?author2 .
}
```

Some remarks are in order. 1) This example query is rather crude and would have to be expanded to handle multiple authorship. 2) Some queries are easier to perform with one or more intermediate extraction queries. 3) Although extraction queries are obviously not as powerful as a dedicated program or full-fledged reasoner, they are often sufficient and much faster to implement.[2]

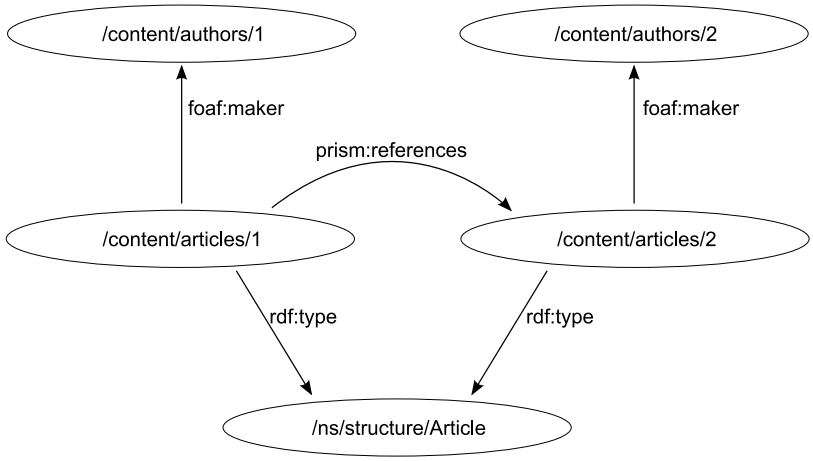

Fig. 1. Citation relation in the IngentaConnect MetaStore [22] with base URI `http://metastore.ingentaconnect.com`

2.2 From secondary graph to SNA format

Once a secondary graph has been obtained, it can be studied. There exist several projects for visualizing and exploring RDF and FOAF data, such as FOAF Explorer,[3] RDF-Gravity[4] and Visual Browser.[5] These tools, however, generally do not provide SNA measures like centrality and clustering. Moreover, they generally do not scale to very large graphs.

Thus, while not strictly necessary, this step ensures compatibility with other SNA efforts and permits techniques that are difficult to perform on plain RDF graphs. We handle these conversions by integrating with pyNetConv, a Python library that can convert to Pajek, NetworkX, CytoScape, GML, ...

[2] Some triplestores, like Jena, also allow custom SPARQL functions.
[3] `http://xml.mfd-consult.dk/foaf/explorer/`
[4] `http://semweb.salzburgresearch.at/apps/rdf-gravity/`
[5] `http://nlp.fi.muni.cz/projekty/visualbrowser/`

3 Unevenness

The distribution of degrees on the Semantic Web is — like many other relations — highly uneven: a small number of nodes has a huge amount of links, while the vast majority has very few. How can this unevenness be quantified?

Unevenness or inequality has been studied extensively in econometrics and informetrics. Since not all existing measures satisfy all necessary requirements [1, 14], we will limit the present discussion to two methods, using the following array as an example: X = (1, 3, 4, 7, 10, 15). These numbers could e.g. express the distribution of wealth or the distribution of degrees for a set of nodes. Clearly, there is some unevenness, but how much exactly?

The *Lorenz curve* [18] is a graphical representation of unevenness. First, we determine the relative amounts:

$$a_i = \frac{x_i}{\sum_{j=1}^{N} x_j} \tag{1}$$

resulting in $(\frac{1}{40}, \frac{3}{40}, \frac{1}{10}, \frac{7}{40}, \frac{1}{4}, \frac{3}{8})$. The horizontal axis of the Lorenz curve has the points i/N $(i = 1, 2, \ldots, N)$. The vertical axis of the Lorenz curve has their cumulative fraction: $a_1 + a_2 + \ldots + a_i$. We thus construct the Lorenz curve (Fig. 2). The diagonal line represents the case of perfect evenness. The further the curve is removed from the diagonal, the greater the unevenness. Note that we have ranked our numbers in increasing order, resulting in a convex Lorenz curve. The concave Lorenz curve results from ranking in decreasing order and is completely equivalent. Complete unevenness — one person has everything, and the rest nothing — would be represented as a curve following the bottom and the right side of the plot.

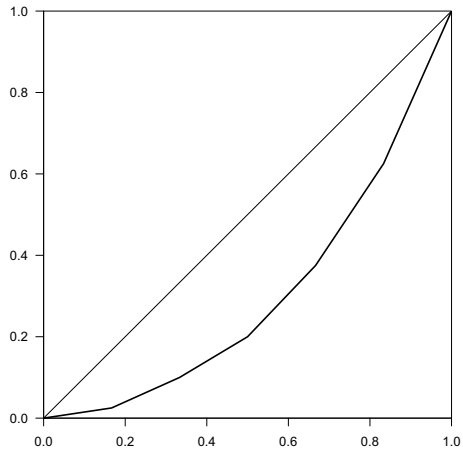

Fig. 2. Convex Lorenz curve of the array (1, 3, 4, 7, 10, 15)

Suppose we want to express this unevenness in a number. A good measure is the *Gini evenness index* G' [25], originally devised to characterize the distribution of wealth and poverty [16],

$$G'(X) = \frac{2}{\mu N^2}\left(\sum_{j=1}^{N}(N+1-j)x_j\right) - \frac{1}{N} \tag{2}$$

with x_j ranked in increasing order and μ the mean of the set x_j. $G' = 2 \times$ the area under the convex Lorenz curve.

Lorenz curves determine a *partial order*: if one convex Lorenz curve is completely below another, then the former expresses less evenness than the latter. It should be stressed that Lorenz curves may 'overlap' or cross each other. In these cases, no order can be determined [25].

4 Example: Agrippa

For this example, we use data from the *Agrippa* database, the catalogue and database of the Archive and Museum of Flemish Cultural Life (AMVC Letterenhuis, Antwerp). Agrippa contains a wealth of information about both the archived materials and the socio-cultural actors that have created them. The RDF version uses existing ontologies like FOAF and Dublin Core, where applicable. The graph is stored in a Jena triplestore and made available via the SPARQL protocol [11] using Joseki.[6] Through this protocol, SPARQL queries can be submitted to a centralized server.

Many secondary graphs can be derived. The following, for instance, constructs a bipartite graph of persons and their affiliations to organizations.

```
PREFIX agrippa: <http://anet.ua.ac.be/agrippa#>
CONSTRUCT { ?person agrippa:affiliatedWith ?org }
WHERE {
  ?aff agrippa:hasAffiliator ?org .
  ?aff agrippa:hasAffiliatee ?person .
}
```

Agrippa also contains information about 237,062 letters. We construct a simple graph that links author(s) and recipient(s) of each letter:

```
PREFIX agrippa: <http://anet.ua.ac.be/agrippa#>
CONSTRUCT { ?sender <urn:agrext#writesLetterTo> ?recipient }
WHERE {
  ?context agrippa:hasLetterWriter ?sender .
  ?context agrippa:hasRecipient ?recipient .
}
```

[6] `http://joseki.sourceforge.net`

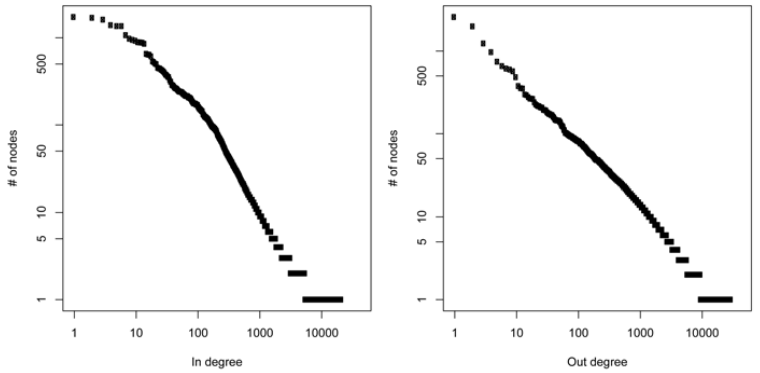

Fig. 3. Zipf distribution for in-degree and out-degree

We will take this author-recipient graph ($N = 40,914$) as an example. Each node is connected by 5.08 links on average, but the actual in- and out-degree follow a Zipf distribution (Fig. 3). Apart from degree centrality (DC), we also consider the following two centrality measures [26]:

Betweenness centrality (BTC): characterizes the importance of a given node for establishing short pathways between other nodes.

Closeness centrality (CC): characterizes how fast other nodes can be reached from a given node.

Comparing the Lorenz curves of the three centrality measures reveals a remarkably diversified picture, shown in Fig. 4. BTC is clearly more uneven than the other two. In spite of the initial appearance, no order can be determined between DC and CC, since the curves overlap slightly at the bottom (recall that the Lorenz curve imposes only a partial order). The Gini evenness indices are: $G'(\mathrm{BTC}) = 0.02 < G'(\mathrm{DC}) = 0.25 < G'(\mathrm{CC}) = 0.98$.

As a tentative explanation, we suggest that these differences may be due to the small-world effect [21, 27]. Even marginal nodes are relatively close to all others, accounting for minimal differences in closeness. Indeed, the length of the diameter — the longest shortest path — is only 11 and the average shortest path length only 3.85! The graph is not fully connected, but the main component ($N = 40,303$) accounts for the vast majority of nodes. The core of the main component is the Largest Strongly Connected Component or LSCC ($N = 9,723$), a component in which any node can be reached (obeying the direction of the links).[7] The LSCC itself has a nucleus of *hubs* [10], nodes with extremely high DC, through which almost all other shortest paths pass. This

[7] As a whole, the graph fits the bow-tie model [6, 9], previously devised for link structure on the World Wide Web.

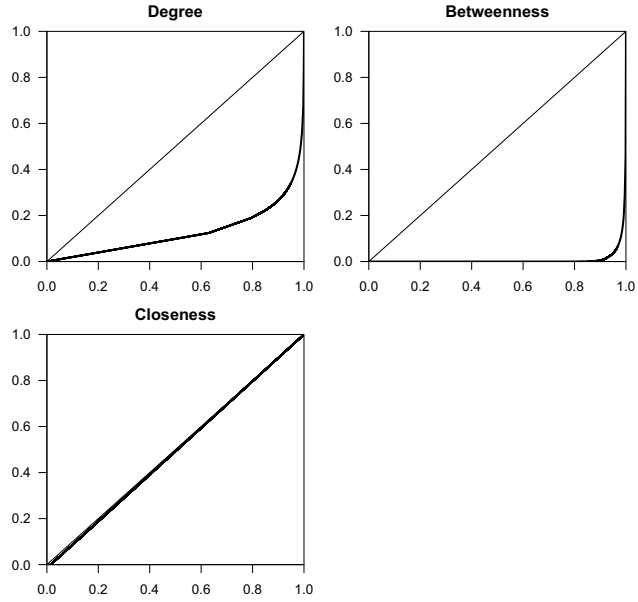

Fig. 4. Lorenz curves for degree, betweenness and closeness centrality

increases closeness for the network as a whole and brings about a very uneven BTC distribution.

5 Conclusions

We have shown how SPARQL can be used in processing social Semantic Web data in a simple two-step methodology, converting the primary graph to a better suited secondary graph. While SPARQL is obviously less powerful than a 'real' reasoning engine or a dedicated program, it is often sufficient and may well prove simpler and faster to implement. RDF tools are generally not geared towards SNA, although Flink [19] incorporates some basic SNA statistics. Generally, conversion to other formats is recommendable but, luckily, straightforward.

The Lorenz curve and the Gini evenness index G' are two excellent methods for studying unevenness. Taking Agrippa as a concrete example, it can be seen that unevenness measures may confirm or enforce hypotheses regarding the network topology. In the example discussed, the massive difference between BTC and CC distribution confirms the small-world hypothesis and reveals the topology of the graph with a small nucleus, through which most other paths must pass.

Most of these results, such as the establishment of the small-world effect, could have been achieved without studying the unevenness of network properties. Consequently, the current paper should be regarded as a first step: it illustrates

how unevenness measures can be used to achieve similar results as existing, well-established methods. In future research, we hope to expand upon these results by studying a greater variety of network properties and (social) networks, including different classes of small-world networks [2].

Acknowledgements: I thank prof. Richard Philips for providing access to the Agrippa dataset and the anonymous reviewers for useful comments on an earlier version.

References

1. Allison, P.D.: Measures of inequality. *American Sociological Review*, 43(6), 865–880 (1978)
2. Amaral, L. A., Scala, A., Barthelemy, M., Stanley, H. E.: Classes of small-world networks. *PNAS*, 97(21), 11149–11152 (2000)
3. Bachlechner, D., Strang, T.: Is the Semantic Web a small world? In: *Proceedings, Second International Conference on Internet Technologies and Applications (ITA 07)*, `http://elib.dlr.de/47899/` (2007)
4. Becket, D.: New syntaxes for RDF. In: *WWW 2004*, `http://www.dajobe.org/2003/11/new-syntaxes-rdf/paper.html` (2004)
5. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American Magazine*, 284(5), 34–43 (2001)
6. Björneborn, L.: *Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach*. PhD thesis, RSLIS, Copenhague (2004)
7. Bollen, J., Rodriguez, M. A., Van de Sompel, H., Balakireva, L. L., Hagberg, A.: The largest scholarly semantic network... ever. In: *Proceedings of the 16th International Conference on the World Wide Web*, ACM Press, 1247–1248 (2007)
8. Brickley, D., Miller, L.: FOAF Vocabulary Specification 0.91. Namespace Document 2 November 2007 – OpenID Edition, `http://xmlns.com/foaf/spec/` (2007)
9. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the Web. *Computer Networks*, 33(1–6), 309–320 (2000)
10. Christensen, C., Albert, R.: Using graph concepts to understand the organization of complex systems. `http://arxiv.org/abs/q-bio/0609036` (2006)
11. Clark, K. G., Feigenbaum, L., Torres, E.: SPARQL Protocol for RDF. W3C Recommendation 15 January 2008, `http://www.w3.org/TR/rdf-sparql-protocol/` (2008)
12. Ding, L., Finin, T., Joshi, A.: Analyzing social networks on the Semantic Web. *IEEE Intelligent Systems*, 1(9) (2005)
13. Ding, L., Finin, T.: Characterizing the Semantic Web on the web. In: *Proceedings of the International Semantic Web Conference*, Springer (2006)
14. Egghe, L., Rousseau, R.: Transfer principles and a classification of concentration measures. *Journal of the American Society for Information Science*, 42(7), 479–489 (1991)
15. Gil, R., García, R.: Measuring the Semantic Web. In: *Advances in Metadata Research, Proceedings of MTSR '05*, Rinton Press (2006)
16. Gini, C.: Il diverso accrescimento delle classi sociali e la concentrazione della richezza. *Giornale degli Economisti*, 11(37) (1909)
17. Lee, R.: Scalability report on triple store applications. `http://simile.mit.edu/reports/stores/` (2004)

18. Lorenz, M. O.: Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70), 209–219 (1905)
19. Mika, P.: Flink: Semantic web technology for the extraction and analysis of social networks. *Journal of Web Semantics*, 3(2), p. 211–223 (2005)
20. Milgram, S.: The small world problem. *Psychology Today*, 2(1), 60–67 (1967)
21. Newman, M. E. J.: Models of the small world. *Journal of Statistical Physics*, 101(3), 819–841 (2000)
22. Portwin, K., Parvatikar, P.: Building and managing a massive triple store: An experience report. *XTech 2006: "Building Web 2.0"*, `http://2006.xtech.org/schedule/paper/18/` (2006)
23. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF. W3C Recommendation 15 January 2008, `http://www.w3.org/TR/rdf-sparql-query/` (2008)
24. Rodriguez, M. A., Bollen, J., Van de Sompel, H.: A practical ontology for the large-scale modeling of scholarly artifacts and their usage. In: *JCDL '07. Proceedings of the 2007 Conference on Digital Libraries*, ACM Press, 278–287 (2007)
25. Rousseau, R.: Lorenz curves determine partial orders for comparing network structures. *Critical Events in Evolving Networks (CREEN) Workshop, Brussels* (2007)
26. Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences 8, Cambridge University Press (1994)
27. Watts, D. J., Strogatz, S. H.: Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440–442 (1998)